LEVERAGING AUTHENTIC MEDIA TO DESIGN SCALABLE FOREIGN LANGUAGE LEARNING SYSTEMS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Gabriel Rubow Culbertson

August 2018

LEVERAGING AUTHENTIC MEDIA TO DESIGN SCALABLE FOREIGN LANGUAGE LEARNING SYSTEMS

Gabriel Rubow Culbertson, Ph.D.

Cornell University 2018

Many people want to learn a foreign language, but issues of time, convenience and cost mean that classes are often insufficient for learners, and increasingly learners are supplementing or replacing their classroom learning with online learning. The majority of existing online language learning systems use the traditional grammar-translation approach (i.e. a focus on grammar and translation skills) to language learning. This approach has inherent scale limitations because learning materials must be carefully designed by experts, often requires that learners engage with topics that are irrelevant to their goals and interests, and lacks contextual information that is important for language skill. However, drawing from communicative language approaches (i.e. approaches that focus on meaning rather than correctness, and assess learners based on the activities that they can engage in rather than their knowledge of rules and vocabulary), we can reimagine the design of online language learning to overcome these challenges. Through discussion of three projects, I show that by leveraging readily available native-speaker media, automation, and communicative-learning approaches, we can use (i) learner activity from even novice learners to annotate learning materials (e.g. captions and phonetic readings), (ii) freely available videos and speech recognition to enable contextualized learning practice, and (iii) videos and captions to generate automated learning assessments that capture general proficiency rather than specific vocabulary and grammar knowledge.

This work makes contributions in areas of design, language education, language research methodology, and language learning theory. In design, this work shows how we can build effective language learning experiences around existing materials. In language education, this work generated new learning systems which have been used by independent learners and in classrooms. In language research methodology, this work contributes a new way for researchers to assess learner proficiency using a quick test generated from existing materials. Finally, in language learning theory, this work shows a paradigm shift from the grammar-translation approach to a communicative approach in language system design.

BIOGRAPHICAL SKETCH

Gabriel Culbertson completed his Bachelor's of Science in Mechanical Engineering at Purdue University in 2014. During his undergraduate study, Gabriel spent a year and a half in China. During this time, he studied at Shanghai Jiaotong University, completed an engineering internship in Wuhan and began his journey into design for language learning. The time in China eventually served as inspiration for Gabriel's Ph.D. thesis work at Cornell University on design for foreign language learning.

This document is dedicated to my loving wife Yixuan.

ACKNOWLEDGEMENTS

To my co-advisers Malte Jung and Erik Andersen. From each of you I learned new ways to design, write and think. Both of you challenged me and helped me grow in so many ways.

To my committee member Morten Christiansen. It was a pleasure to learn about and discuss language learning with you.

To my wife Yixuan Li who was the first to test everything I made.

To Solace Shen, who was always there to chat and give me honest feedback.

To Hamish Tennent who taught me about design and presentation.

To all the members of the Robots in Groups Lab. You all offered me an endless supply of support and feedback.

To all the members of Sue Fussell's lab. From you all I gained new perspectives on research, culture and communication.

To all the students who helped me develop language learning systems, especially Andrew Jiang, Shiyu Wang, Liane Longpre, Karrie Shi, and Eileen Liu.

To my parents who showed me how to be curious and seek adventure.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Learning a second language is a long and challenging process. For example, a study of students in the United States learning English as a second language estimated attaining oral proficiency in English takes 3 to 5 years [44]. Beyond this, academic-level language proficiency is estimated to take 5-7 years. The students in this study lived in the United States and thus had the benefit of constant passive exposure to English and plentiful opportunities to practice. Other work has suggested that for native speakers learning their first language, proficiency does not peak until as many as 30 years [49]. While we lack definitive durations for learners in other environments, we would expect learners in classrooms where language experiences are limited to just a few hours a week to need significantly longer to reach even basic language proficiency. Given the extent of commitment needed to learn a foreign language, few students are able to succeed through classes alone. Even if a student takes four years of language classes, there may simply not be enough time to achieve proficiency. Furthermore, mismatches between the skill and interests of individual students with course content can reduce the effectiveness of learning in classroom settings. Beyond issues of time and personalization, classes may not even be an option learners who are already working or cannot afford classes. Together, these factors mean that classes are insufficient to support language learning, and many learners will need to use alternative learning methods.

Given this challenge, many online systems have been developed to support language learning. With online learning, students can learn when they choose, learn at their own pace and learn without making long-term time commitments. While online learning systems have offered new opportunities for language learning,

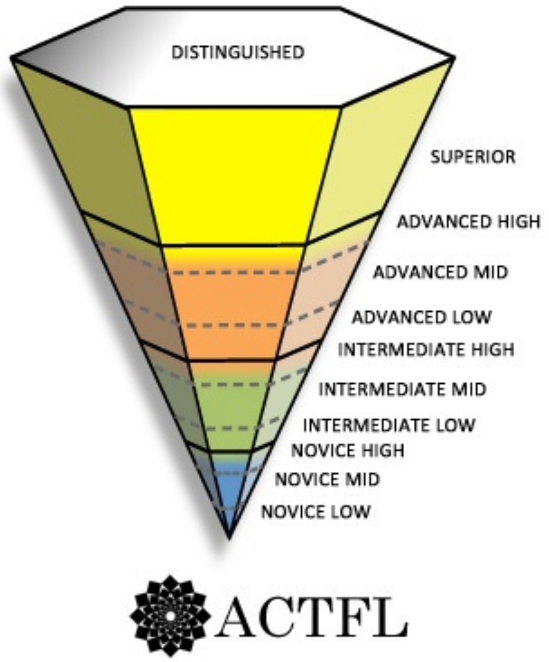

Figure 1.1: In their proficiency guidelines, the American Council for Teaching of Foreign Languages represents proficiency as expanding in scope at each level [2]. This illustrates the increasing effort that learners will need to invest at each subsequent level.

the overwhelming majority of both the commercial and research language learning systems are designed with a focus on grammar and translation skills (grammar-translation approach). In this perspective of language learning, the goal of language education is to learn word lists and grammar rules [79]. Approaching language education in this way has inherent limitations for volume of content, fit of learner goals to learning topics, and contextualization of language knowledge.

First, if we assume learners are unable to learn from content that falls outside of already learned rules, then language content must be carefully developed to fit within the constraints of the learner's ability. Because expert educators or researchers need to develop this content, this approach has limited scalability. We see this limitation reflected in the majority language learning systems.

Level	Kanji	Vocabulary
N5	100	800
N4	300	1500
N3	650	3750
N2	1000	6000
N1	2000	10000

Figure 1.2: The estimated number of Kanji and Vocabulary needed to advance between each level roughly doubles for each level [1].

For example, completing all of Duolingo (a website that teaches language through translation, matching and various other language exercises [1]) in Japanese is reported to only help learners reach JLPT N5 (e.g. "The ability to understand some basic Japanese." [89] in a widely adopted Japanese assessment) or A1/A2 (basic ability to communicate) in the Common European Framework [47]. Furthermore, it would likely be very difficult to develop the content needed to go beyond this point. The scope of language rapidly expands as learners increase in level (e.g. ACTFL figure). As a more concrete illustration of this effect, we can observe that the amount of vocabulary needed to reach the next level in the JLPT roughly doubles at each level (Figure 1.2). Or, in English, we can observe that the 100 most common words make up around 50% of the language, but to reach 95% coverage, a vocabulary of around 50,000 words is needed [67]. Thus it requires significantly more effort to design a system to help a learner make progress at higher learner levels. While researchers have investigated novel language learning systems, the narrative in language learning system research is most frequently centered around convenience, efficiency and learner motivation. Content depth was not discussed or considered in any of the 53 language learning system papers reviewed in the ACM library. Together this means that learners can find many opportunities to learn in the initial stages of a language, but will quickly lose support as they progress.

[1] duolingo.com

Second, when language curricula or systems focus on grammar rules, it becomes difficult to adapt learning to the specific goals and interests of individual learners. While early in the learning process content is generally relevant to all learners, as learners advance and content becomes more specific, it becomes increasingly difficult to identify topics that will be relevant to all learners. Consider the Genki textbook [8]: the first chapter of the first volume covers greetings and numbers (relevant almost all learners) whereas the first chapter of the second volume covers applying for part-time jobs (relevant to far fewer learners). Furthermore, if the topic of particular chapter or learning module is uninteresting to a learner, it is difficult to skip. For example, in the Chinese textbook New Practical Chinese Reader [100], there is a chapter on the Beijing Opera. This topic may be irrelevant to many learners, but they would need to complete the chapter anyway because grammar rules from the chapter will appear later. We find similar issues in many learning websites (e.g. Duolingo[2], Mango Languages[3], Rosetta Stone[4]) where learning modules need to be completed in a specific order. Given the extensive effort needed to develop language content, once content has been developed for a specific progression of grammar rules, it would likely be very challenging to adjust that content for each individual learner. The grammar-translation approach prioritizes abstract rules over relevant topics, so learners may often find themselves engaging in irrelevant material.

Finally, the grammar-translation approach largely ignores context in language. The way that we express ideas and understand expressions is highly dependent on context. For example, if someone had their hand full and wanted to request help from a passerby, "Would you mind opening the door?'' and "I wish the door

[2]duolingo.com
[3]mangolanguages.com
[4]rosettastone.com

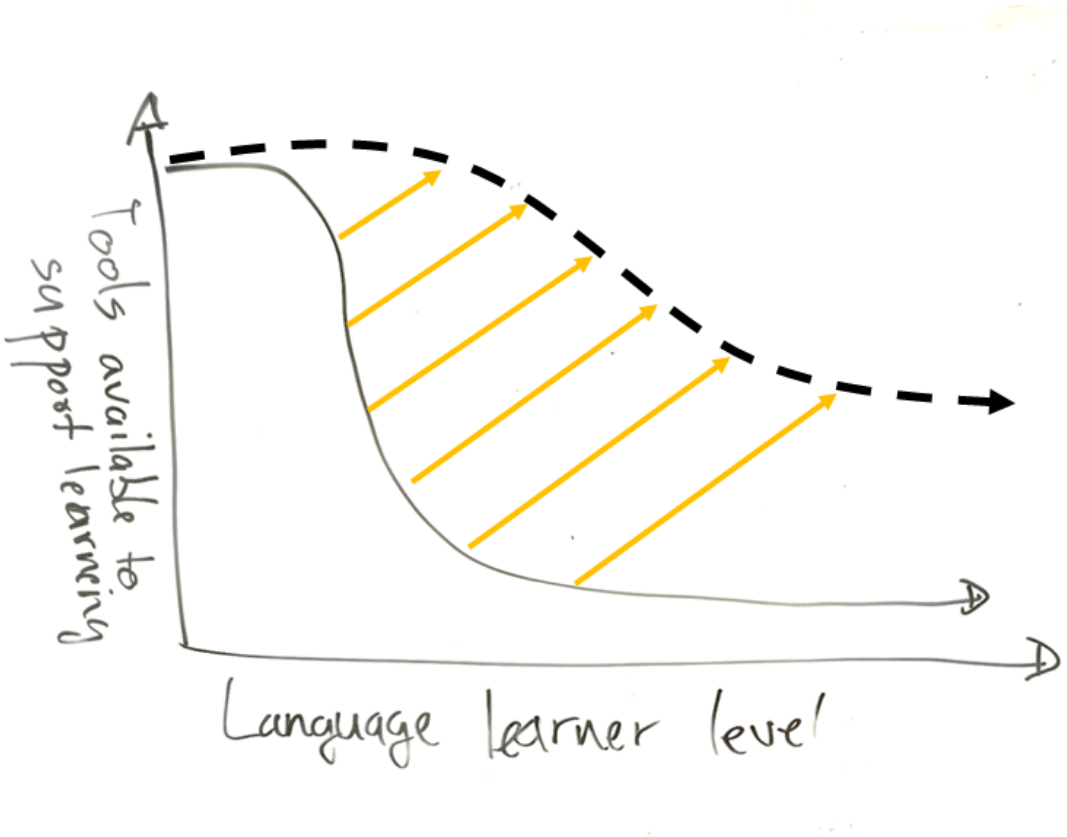

Figure 1.3: This figure illustrates my understanding of a key challenge in language system design. From my perspective, although many tools are available for learners initially, currently few tools are available to learners at advanced levels (as indicated by the solid line). The goal of my work is to rethink design of language learning systems to utilize authentic resources in order to scale to any language learner level (expanded support for learners at higher levels is indicated by the dashed line).

to be opened by you'' could both be used to express the request and are both grammatically correct, but the first is much more likely to be understood and well received [79]. Most existing language learning systems lack any meaningful context (e.g. Duolingo and Rosetta Stone present only single sentences or simple images). Using just the knowledge from one of these systems, the learner does not have any experience to refer to when sifting through innumerable ways an idea

can be expressed. Textbooks may offer more context in the form of dialogues, but these scenarios are sparse relative to the countless diverse situations learners will encounter. Pragmatic skill is important to functioning in a culture (e.g. [56]), but, in the traditional language learning paradigm, these skills are not considered part of language proficiency. Ultimately lack of attention to the connection between context and language content will limit how far a learner can develop their proficiency using these systems.

Given challenges with the grammar-translation approach, in language education research the dominant paradigm has moved away from grammar-translation, toward communicative methods. These methods focus on meaning rather than correctness, and assess learners by the activities that they can engage in rather than the rules that the know or the size of their vocabularies [79]. I suggest that by adopting this paradigm in HCI, we can reimagine the design of language learning tools to address challenges in online language learning.

A widely adopted approach to classroom learning with communicative methods is to use authentic resources (i.e. materials designed for native speakers such books, videos, websites, etc) [74] as instructional materials. These materials are designed to communicate ideas between native speakers and therefore are often accurate reflections of how native speakers express ideas in a given context. In online learning, we have access to countless freely available websites, podcasts and videos in almost every language. Other forms of media such as novels and games can also easily be purchased internationally through online platforms such at Amazon or Steam. However, the usefulness of these materials is limited by the lack of supporting tools or systems. In classroom contexts, instructors can create activities around native speaker materials (e.g. answer comprehension questions)

and offer explanations for unusual and difficult language. Outside of classroom contexts, learners using natives speaker materials are mostly limited to tedious use of dictionaries or machine translators [24]. While authentic resources are useful for supporting communicative learning approaches, it is currently very difficult to use these resources outside of classroom contexts.

In this thesis, I show that by supporting authentic-resource-driven (learning using authentic materials as the primary content) study with digital tools, we can reimagine the design of education technology to address challenges of content volume, specificity and contextualization. With the vast number of resources available, learners can choose media that are interesting and relevant to their interests and goals. This also helps learners to avoid challenges with the overwhelming amounts of new content at later language levels by focusing learners on activities they intend to engage in. Furthermore, by leveraging automation (such as speech-recognition and crowdsourcing), we can help learners make sense of materials and practice language effectively. By designing systems to support learning through authentic materials, rather than simply teaching abstract grammar and vocabulary, we can help learners improve their language proficiency in countless detailed and interesting contexts.

Thesis statement: By leveraging readily available native-speaker media, automation, and communicative-learning approaches, we can design scalable, personalized and contextualized learning systems. In particular, we can use (i) learner activity from even novice learners to annotate learning materials (e.g. captions and phonetic readings), (ii) freely available videos and speech recognition to enable contextualized learning practice, and (iii) videos and captions to generate automated learning assessments that capture general proficiency rather than spe-

cific vocabulary and grammar knowledge.

In the following chapters, I describe three projects which illustrate how to design using this apporoach.

In Chapter 3, I discuss a method for resource collection and annotation with crowds of learners. As I have discussed, freely available foreign language videos are abundant and useful resources. However, the videos on their own often are insufficient for learning. One of the most helpful ways to supplement video materials is through foreign language captions, but these captions are often unavailable. In this project, I show how we can leverage imperfect automatic transcripts and learner activity to create accurate captions for foreign language videos. By using a machine generated transcript as a starting point and alternative translations generated by the translator to reduce the answer space, learners of any level can contribute to the caption creation process and learn at the same time.

In Chapter 4, I discuss the design of learning experiences that can be used with any native speaker resource while encouraging learners to practice speaking skills in context. Even with useful resources such as captioned videos, learners can benefit more from these resources if they practice actively (e.g. speaking aloud) than if they engage passively (e.g. reading captions). In this project I designed a system for learning from any foreign language video by speaking aloud and providing feedback through automated speech recognition. When using the system, learners can select any video from YouTube or their hard drive and practice speaking in the context of the video. Learners can repeat phrases from the video, or roleplay as one of the characters. The system displays a transcript of the learner's speech and an automatic translation. This way the learner can practice their speaking skills along with speech from native speakers and learn using materials that are

interesting and relevant.

In Chapter 5, I discuss the design of learning measures that can be automatically created, automatically scored, and broadly reflect a learner's proficiency. In this project, I show that by using audio clips that can be automatically extracted from videos with captions and a verbal recall task, we can obtain a strong measure of a learner's comprehension ability and general proficiency. In the test, learners listen to audio clips with a controlled length and write down as much of the utterance as they can remember. Using the captions as the correct answer, the written utterances can easily be scored for how many correct words were written down. Finally, I show that this measure is not only correlated with a general comprehension proficiency test, but is even a better predictor of a learner's ability to translate heard utterances than a standardized listening comprehension test.

CHAPTER 2

RELATED WORK

Previous research has highlighted difficulties in foreign language education. Many learners are unable to communicate effectively even after many years of learning. In this chapter, I describe previous work showing that language learning continues to be a be hard for students, and that students that are unsuccessful in language learning may see negative impacts in their lives. I then provide an overview of first language acquisition theories and show how these theories are reflected in classroom language education, especially in classroom approaches using native speaker materials. Communicative and task-based curricula using real scenarios and native speaker materials have been widely adopted and been shown to be effective. I then provide a summary of online foreign language education and discuss some of the essential limitations in existing online education. I suggest that these limitations can be overcome by utilizing native speaker materials as they have in classrooms, but using native speaker materials in online contexts brings a new set of challenges. Finally, I describe methods from HCI that can be leveraged in online learning with native speaker materials that suggest ways to overcome these challenges.

2.1 Challenges with foreign language competence

Language learning is hard and many people are ultimately unsuccessful. Even of those who reach higher levels of proficiency, many continue to have challenges communicating effectively in their learned language. For example, in an interview study, Yuan et al. [102] showed that non-native English speakers in universities found it difficult to interact with native speakers in part because of lack of com-

mon ground and unfamiliarity with informal and idiomatic English. These Chinese students began learning as early as elementary school, but still struggled in out-of-classroom contexts. Communication styles differ between native and non-native speakers [66, 73] and that these differences often impair successful communication. For example, in a lab study, Wong [98] showed that native Mandarin speakers had difficulty with repairing misunderstandings in face-to-face communication. Some studies have shown that even an accent may trigger certain biases in native speakers [58]. For example, native speakers are more likely to re-interpret implausible utterances as more plausible utterances when the speaker has an accent [38], and the same facts are viewed as less reliable when the speaker has an accent [65].

Furthermore, even subtle nuances of word choice and tone can impact interactions. These nuances are collectively known as pragmatics: "the study of language from the point of view of users, especially of the choices they make, the constraints they encounter in using language in social interaction and the effects their use of language has on other participants in the act of communication" [22]. Pragmatics is an essential part of communication. Kasper [56] provided the following example: "feed the cat", "can/could/would you feed the cat?", and "the cat's complaining" all communicate the same request, but may have very different impacts on the relationship between the requester and interlocutor. In a study of interpretation of dialogues, Bouton [12] showed that in 27% of cases, non-native speakers with high levels of linguistic competence interpreted the meaning of indirect statements differently from native speakers [56]. Furthermore, the inability to express requests effectively can disadvantage non-native speakers. Bardovi and Hartford [9] showed that in academic advising, non-native speakers were less likely to have their requests fulfilled because of a lack of pragmatic competence.

Although some aspects of pragmatic competence learning can be transferred from a learner's first language, when norms differ across cultures, this transfer can hurt rather than help learners. In a study of Iranian learners of English, [5] showed that learners with higher proficiency made more pragmatic errors than low-proficiency learners in making refusals because higher proficiency learners tended to transfer more pragmatic knowledge. Because pragmatics differ from language to language, it is important to learn them. This transfer of pragmatic knowledge is nicely illustrated through a classroom exchange between two students that was observed by Hamid and Naeimi [5]. In this exchange, an upper-intermediate Iranian learner of English who plays a teacher responds to a classmate playing a student: "I have been teaching for many years, and I have experienced many paths. I think it's the best way" using formulaic Iranian refusals whereas a native English speaker responded with "Thanks for your suggestion, but we're following a very strict curriculum" [5]. In sum, if a learner wishes to communicate effectively in a second language, pragmatic competence cannot directly transfer from the learner's native language.

To better understand learner perceptions of obstacles, I surveyed undergraduate language learners at Cornell University [24]. Participants were asked to report their biggest challenge in learning a language and results were coded. When asked about the greatest challenge that learners faced, 41% of learners learners reported challenges related to oral competence: 14% said conversation (e.g. Speaking in conversation as a native would), 8% said limited interaction with native speakers (e.g. it was very hard to find other speakers to practice with), 8% said nuances in language (e.g. Remembering the correct use of certain words even though they have the same meaning), 8% said native-like pronunciation (e.g. Accents to sound natural), and 3% said listening (e.g. having to extract what a speaker was saying).

Some learners also indicated difficulty finding resources for improving oral proficiency: "one of the most significant challenges was accessing learning resources that allowed me to practice speaking and listening''. A significant portion of respondents report challenges that align well with the challenges previously reported in research. Learners hope to communicate like native speakers, but struggle to reach that point.

Language competence is important for navigating everyday scenarios and poor competence can put learners at a disadvantage. However, reaching advanced competence requires careful attention to not only language structure and vocabulary, but also nuances such as pragmatics and pronunciation. In the next section, to investigate how such a complex skill can be learned, I describe prior research in first language acquisition and explain how this knowledge links to foreign language education.

2.2 Understanding first language acquisition

Traditionally language classrooms and online tools have focused on isolated aspects of language skill such as pronunciation, words and grammar. This trend continues to be seen in the majority language system research. For example, during the last 10 years in the ACM community, the vast majority of language systems have focused on one of pronunciation training, abstract grammar learning, or vocabulary memorization. However, overwhelmingly research shows that our knowledge of each of these language aspects is not isolated, and is connected to verbal and physical contexts. In this section I will describe how each of these skills (pronunciation, vocabulary and grammar) are contextualized and cannot be viewed in

isolation.

Pronunciation tools often train individual sounds or words. However, in real contexts, pronunciation of individual words will change based on the surrounding words and overall intonation of the utterance. For example, studies of coarticulation have shown that when speaking, before the speaker finishes one word, they will begin to form the sounds for the text word. This means that the beginnings are and ends of words will lack the clear boundaries that they have when practicing them in isolation [48]. This dependence on context extends to tones in tonal languages as well. For example, the tones of individual characters in Chinese can change depending on the other characters in a word, position in the sentence, and overall sentence intonation [16]. Furthermore, we use more than sound alone to make decisions about pronunciation. For example, McGurk [68] has shown, overlaying video footage of a person saying "ga" over an audio clip of a person saying "ba" results in a hearer interpreting the utterance as "da". The effect has been replicated in other languages, such as Japanese, as well [82]. These examples show that pronunciation is dependent on context, and knowledge of pronunciation cannot be fully learned in isolation.

The majority of research into developing language learning systems has focused on developing systems to support vocabulary learning. However, again, our knowledge of vocabulary is connected to context. First, words are not simply links to abstract concepts. For example, a study Hayakawa and Keysar [50] shows that metal images created by hearing foreign language words are less clear than mental images found by first language words. They showed that when asked which object of three was most dissimilar in shape, they were less accurate in the task if the words were in a foreign language. This means, simply knowing the translation of

a word is insufficient to lead us to react to that word in the same way a native speaker would. Our knowledge of words is also linked to physical context. It has been shown that recalling words in the same context that they are learned leads to better recall accuracy [86]. In the study, participants learned a list of words and either recalled them in the same physical context or different physical context, and recalled more words when in the same physical context. This shows some part of the learners knowledge about the context has been linked to the words being learned.

Knowledge of words also encompasses knowledge about which words are likely to appear around a given word. Studies of word associations (e.g. [69]) have shown that native speakers and non-native speakers respond with different sets of words when prompted with a word (e.g. prompt: "game'', followup: "win'', "play'', "ball'' etc.). This shows that word associations are not simply properties of physical relationships between objects, but also of the contexts words are learned in. Finally, multi-word-chunks have been shown to be an important way that we process language [19]. For example, reading the garden path sentence "The old man the boat.'' we are much more likely to group together "The old man'' (leading to a grammatically incorrect interpretation of the sentence) because we automatically group together the highly frequent combination of "the old man''. These examples show our sensitivity to how words appear in combination and how this knowledge becomes part of our language skill.

Finally, a top-down, grammar-first approach to language processing has been shown to insufficiently describe language processing. Work by Christiansen and Chater has shown that there is a fundamental constraint on first language acquisition: language is learned verbally, but working memory can only hold a very small

amount of audio [20]. This constraint is known as the "now-or-never bottleneck". It is not possible for learners to remember more than a few fractions of a second of audio at any given time. Studies have shown that audio memory only lasts about 100 milliseconds [20]. To gain an intuitive understanding of this limitation, the reader can listen to a few seconds of audio in an unknown foreign language and try to mentally replay the audio afterwards. Similarly, we have difficulties recalling random strings of words that are too long. Without employing memorization strategies, we might have difficulty recalling even the 7 digits of a phone number.

Given that working memory is so quickly overwritten, our minds must process and combine information rapidly and incrementally. This is in contrast to a traditional view of grammar that would suggest we process sentences as a whole using a tree-like grammar. As a counterexample to this view, consider work on garden path sentences. For example, "The old man the boat.". In this sentence, listeners will most likely greedily chunk the first three words ("[The old man]"). After chunking these words together, most people will be unable repair the their understanding to the correct form ("[The old (N)] [man (V)] [the boat (DO)]") by reconsidering the entire sentence. Fierra et al. showed that most people do not repair their understanding when encountering garden path sentences. In their work, most participants sustained misconceptions after reading these types of sentences. For example, after reading the sentence "While Anna dressed the baby spit up on the bed." most participants incorrectly believed that Anna dressed the baby (whereas the grammatically correct interpretation is that [While Anna dressed] [the baby spit up on the bed]). This evidence supports that idea that language is processed forward, in sequence, and information that has already been processed is lost.

Because grammar cannot fully account for the way we process language, other information must be used to make decisions about the boundaries of a chunk of information. Language research has shown that we use a variety of cues from many different sources. One important factor is semantics. If we changed the earlier garden path sentence to "While Anna ate the baby spit up on the bed", participants are unlikely to retain the belief that "Anna ate the baby" although grammatically this should be equally as likely as the belief "Anna dressed the baby". In another study comparing interpretations of plausible and implausible garden path sentences, Christianson et al. [21] showed that semantically implausible sentences are misinterpreted at approximately the same rate as non-garden-path sentences, but plausible sentences are misinterpreted at a significantly higher rate. Other work has shown that hearing words can prime comprehension of other words [7]. For example, being prompted with the word "telephone" can increase the speed with which participants can recognize "pole" in a lexical decision task. This finding suggests that we use recently heard words to facilitate comprehension of following words. Other work using computer modelling has shown that multiple cues including statistically likely phonetic endings can account for learning word segmentation [18]. They used a model trained on child directed speech to show that word segmentation is learned most effectively when considering multiple factors (e.g. lexical stress, words that appear at the end of sentences, phonetically likely word endings). Together, the diversity of the factors shown to facilitate language language comprehension and recall show that language acquisition is tied contexts of learning, whether those contexts be life experiences (as shown in the semantics example), word collocation (as shown in the lexical decision example) or properties of sound (as shown in segmentation using factors like emphasis and statistically likely word endings).

Learning to combine and simultaneously utilize the many cues which we use to comprehend and produce language is complex. According to Christian and Chater [20], because this information vanishes from our minds so rapidly, any learning must happen immediately and through practice.

> [The] Now-or-Never bottleneck requires that language acquisition be viewed as a type of skill learning, such as learning to drive, juggle, play the violin, or play chess. Such skills appear to be learned through practicing the skill, using online feedback during the practice itself, although the consolidation of learning occurs subsequently (Schmidt & Wrisberg 2004). The challenge of language acquisition is to learn a dazzling sequence of rapid processing operations, rather than conjecturing a correct "linguistic theory." [20]

Language learning is a skill and an abstract concept of language rules is insufficient for language learning. In the same way that reading a book about playing tennis might give you insight into into techniques to try but you would not be able to execute them until actually trying them many times, grammar and other language rules may give insight into strategies for comprehending and producing language but only realtime comprehension and production can lead to language acquisition.

In sum, because sound comes and goes so quickly, learning language requires high-speed combination of incoming information to increasingly consolidated units, and once converted, the original information is lost. This property of language means that we must learn to make instant judgments about qualities of the information we are hearing (where does this word stop?, where does this phrase stop?,

which of the possible meanings does this have?) and we use information from many sources to make these judgments. However, since learning must happen immediately, these sources of information must be available to us at the instant of processing in order for us to learn to use them. In the traditional grammar-translation approach to language learning, these forms of context are often ignored which can impair the acquisition of language skill.

2.3 Language education

Although the research from cognitive psychology discussed in the previous section focused on first language acquisition, we can see many of these ideas supporting changes to foreign language learning pedagogy. The precise extent to which first language acquisition theory applies to second language acquisition has been extensively debated (e.g. [54], others), however, evidence points to some of our understanding of first language acquisition being applicable to foreign language acquisition. The same neurological limitations of short-term memory (the now-or-never bottleneck) exist for learners of first and second languages, and likely overcoming those limitations requires the same basic processes. Moreover, language education that treats language as a situated skill rather than an abstract rule set has been shown to be effective. Language curricula increasingly incorporate inference over explicit rule teaching, situated content over abstract content, and a focus on partial real time comprehension and production over slow or repeated comprehension and production with complete correctness.

The shift in focus from grammaticality and correctness in foreign language education to a focus on comprehension and expression was spurred by work by Stephen

Krashen. Krashen [61] suggested a difference between language "learning" and "acquisition". "Learning" is primarily a conscious process and is connected to grammar or other rules in language learning, whereas "acquisition" is primarily subconscious and focuses on meaning over form. To illustrate this difference, consider a novice pianist who must look at his hands and think about where to place each finger to play the notes. Compare this to an experienced pianist whose hand automatically takes the necessary form without sight or conscious thought. Similarly, a student who has "learned" a language may be able to consciously pick out words one by one to make a complete sentence, whereas a student who has "acquired" language utters complete thoughts without a conscious effort in constructing the utterance. Whether musician or language student, "learned" expressions will leave a negative impression on the listener, and, in education, we should strive for "acquisition".

The negative impacts of even subtle language issues (e.g. accent, pragmatics) reinforces the need for "acquisition" based education. These nuances are situated in specific contexts. For example, even for native speakers, learning to speak and interact with members of the academic community requires a different skill set than interacting players of an online game. In both cases, the vocabulary and pragmatics are learned through participation in the community. In these cases, the necessity of community participation to learn is apparent, but learning through authentic practice is also applicable to general learning. Lave and Wenger [62] suggest that all learning is better viewed as learning to act in specific contexts. In this case, learning is most effective when students learn through taking part in practices rather than learning abstract forms of knowledge.

Traditionally classrooms have lacked authentic practices, because "classroom discourse is highly conventionalized in ways that severely constrain both the quan-

tity and the quality of learners' participation" [10]. However, in more recent approaches (e.g. "communicative language learning''), learners are asked to engage in real communication as a primary source of learning, and are assessed based on the activities that they are able to engage in rather than any specific rules (e.g. "Can do" statements) [1]. Curricula built around these ideas have not only been widely adopted and shown to improve students' ability to act in these contexts, but in some cases lead to greater improvements in traditional grammar tests [80]. A study of a "Task-Based Language Course" in Spanish showed that students who learned through focusing on the real tasks they intended to take part in using Spanish (in this case US-Mexico border patrol) not only were better prepared to engage in their work than learners in general Spanish class, but also gained more grammar knowledge even though this was not taught explicitly [42].

Unfortunately, learners often have limited access to the real contexts the wish to engage in. Because of this, native speaker materials have frequently been used in classrooms as a substitute. For example, in a survey of classroom language students [24], 63% indicated having used native speaker media in their learning. Native speaker materials include speech or text that is nuanced in many of the same ways that real speech is, and feature contexts that are often close to contexts that might be encountered in real practices [39]. In many communicative focused currula, native speaker resources (e.g. novels, radio recordings, video series) form the core of learning content. Such curricula have been shown to be effective for students to develop communicative competence. For example, in classroom study of Japanese students where one group was given textbook-based materials and the other given videos, songs and articles, the authentic resource group showed greater increases in communicative competence [40].

[1] actfl.org

Perhaps the most extensively studied authentic media approach is video watching (or video learning). Videos offer audio from native speakers, culturally relevant visual context and stories that reflect the cultures where they are created. Many learners and educators have found video learning to be highly effective. For example, in a study where one class was given an video-based curriculum and another was given a grammar-based curriculum, it was found that learners were much better at listening comprehension in the video condition and there was no significant difference in grammar acquisition [52]. Learning in this way allows learners to gain cultural knowledge alongside language mechanics such vocabulary and grammar [37, 81, 51]. Furthermore, learning through video can be highly engaging to learners if the learners find the video content interesting [84].

In summary, in classrooms, communicative language approaches have been shown to be effective at preparing students for real tasks. Various types of authentic materials have been used to support communicative curricula, and have been shown to be effective sources of language content for students. However, as previously discussed, classrooms are not a good fit for all learners (considering time, cost and individual differences), and even in classroom settings learners will spend a significant amount of time learning without an instructor present. Therefore, it is useful to consider how we can bring communicative learning approaches to individual learning contexts.

2.4 Language learning tools and systems

Although research in classroom education has shown that communicative methods are effective for developing language competence, classes are not viable for many

learners. After leaving college, classes are expensive for students. For those students attending a university, time is a precious commodity, and those students may not have sufficient time to dedicate to a language. Furthermore, it may be difficult for students to find a class at the right level or at the right time. Given these challenges, there has long been interest in developing resources for independent learning. For example, Pimsleur [2] developed audio tapes that use carefully timed spacing between sentences to help students remember them more easily. Later Rosetta Stone [3] took advantage of personal computers' ability to combine audio and visual cues to enable learners to learn a language without using native language translations. More recently, in Duolingo [4] (Figure 2.1) learners learn by translating sentences and the system provides detailed feedback and structures content to build at a reasonable pace [94]. While these systems have been helpful to learners and have made language learning more broadly accessible, these systems only help students reach very low levels of proficiency. Data on the effectiveness of these systems is scarce, but, for example, in Japanese these systems often lack enough content to help learners reach even the lowest level of the JLPT or A2 (second level of six) in the European measurement system [47]. I argue that these systems fail to extend to higher levels of proficiency because they continue to approach system design with the perspective that all content must be created by researchers or instructors rather than adopting the more open-ended approach that is being used in communicative classrooms.

Rather than focus on building systems with scalable content, system designers often focus on making learning engaging. For example, in order to make Chinese tone learning more engaging, Edge et al. developed Tip Tap Tones to lever-

[2] pimsleur.com
[3] rosettastone.com
[4] duolingo.com

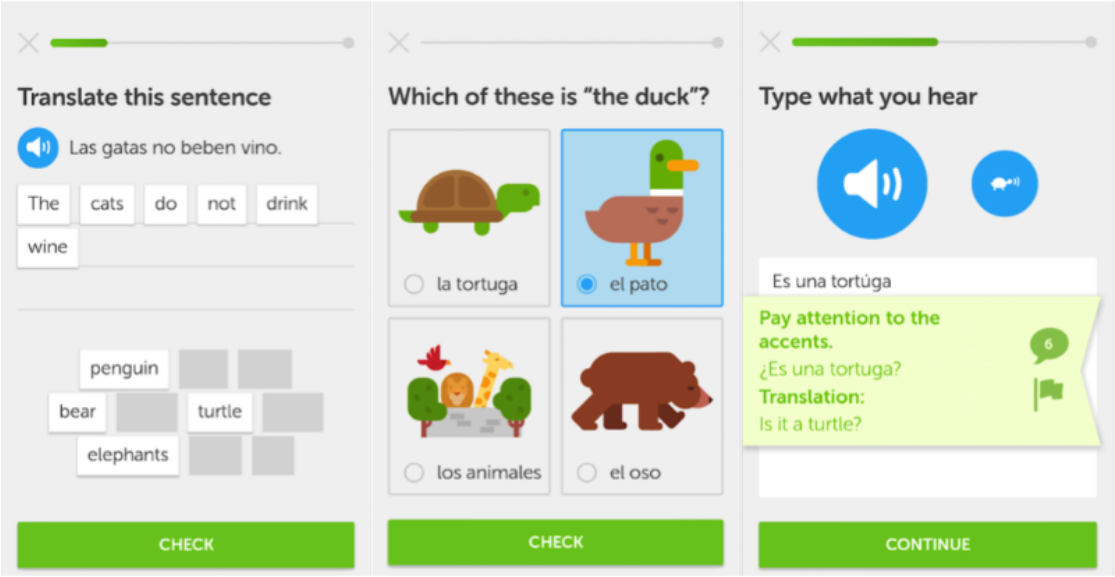

Figure 2.1: Duolingo provides various activities for learners to learn words and sentences (source: https://www.linkedin.com/pulse/getting-started-duolingo-schools-louise-stringer/).

age game-like competition with native speakers to encourage learners [30]. Many adaptive flashcard systems have been developed to support vocabulary learning (e.g. Anki [5], Memrize [6], MemReflex [31]) and maximize retention. Other work has tried to improve vocabulary retention by using a desktop wallpaper to reinforce vocabulary meanings [27] and integrate vocabulary learning into real and simulated environments to give more context to words (e.g. MicroMandarin [32], Influent [7]). All of these systems can benefit learners in some learning tasks, but these systems are unable to help learners with higher-order language skills such as comprehending long texts or utterances, or pragmatics. Thus, while these systems may be helpful in for getting students through early stages of learning and as supplemental activities, again, these system lack the depth needed to help learners

[5] ankiweb.net
[6] memrise.com
[7] playinfluent.com

progress to higher levels of proficiency.

My own early work [23, 26] has helped me to understand the difficulties in building systems that with scalable content. In a previous project, I developed a 3D game to teach language learning in a virtual context. The game combined traditional learning approaches with a situated learning paradigm by integrating a spaced-repetition system within a Japanese learning roleplaying game. To facilitate long-term engagement with the game, we designed "jobs" that were intended to allow a small amount of design effort to generate a large set of highly-scaffolded tasks that grow iteratively and social elements such as chat and shared space. We deployed the game online and had 186 players play the game for an average of 40 minutes.

While the game was successful as a prototype, the tremendous effort that was needed to reach even that point (nearly a year of dedicated work) and the limitations of the game (e.g. the game lacked authentic voice acting, some of the 3D models were purchased so we were limited in available contexts) showed me that this approach could not scale to higher language levels. Even more advanced players played only an average of an hour, and considering the 2000 or more hours that are needed to learn Japanese, this tremendous effort only helped learners through a tiny (less than 0.05% assuming the game is equally as efficient as current methods) of the learning process.

The majority of existing commercial and research language learning systems are severely constrained by the content creation process. Context-rich language learning platforms such as Crystallize [23, 26] or Influent [53] require significant effort to develop not only the language text and scenarios, but the 3D environments to situate that language content. However, even other platforms that have found

ways to reduce the effort needed to produce content (e.g. Duolingo recycles sentences through templating) fail to scale to later language levels, and fail to teach higher level patterns such as speech exchanges (e.g "How are you?", "I'm doing well, how about you?"). Given these challenges, I propose that, rather than study how we can develop content for language learning systems, we should study how to design language learning systems around existing content.

2.5 Facilitating resource annotation

This subsection is an edited version of a subsection that appeared in [24].

Using existing authentic materials as a starting point for design offer nearly unlimited opportunity for depth and scale of language learning opportunities. However, designing around content designed for native speakers brings a unique set of challenges. Raw native speaker materials are often difficult to understand for non-native speakers, but learners are limited in the tools they have to utilize those materials. For example, in a survey of language learners [24], of those who reported using materials designed for native speakers, the most commonly reported method for making sense of the material was a translation system such as Google translate (45%). Other strategies included using subtitles (18%), native speaker friends or family (9%) and continuing to listen despite not understanding (9%). However, transcribing content into Google translate from a video or other resource is tedious, and utilizing experts to annotate materials (e.g. captions) brings the same set of scale limitations as designing new content. However, advances in artificial intelligence and crowdsourcing show new opportunities for material annotation without the need for experts.

While there is some limited work on building systems around existing foreign language media, these systems often continue to be limited in scale. For example, Kovacs and Williams [60] developed an augmented subtitle system that gives participants additional information about individual words in a subtitle and the translation. A commercial system [8] later used a similar approach with commercials and other short video segments. While these systems begin to utilize authentic resources for online language learning, they require experts to annotate the content in order to function and are therefore constrained in the same way as systems requiring new content. A handful of projects do exist to help learners engage with authentic materials. For example, a system has been developed to provide dictionary and grammatical analysis of text in websites [96]. Another system has been developed to provide translations, grammar analysis and text-to-speech in foreign comic books (or manga) [59]. These projects are inspiring examples for how we can design scalable systems around authentic content, but further work is needed to improve the quality and types of available resources, refine interactions that students have with learning systems and assess learning using these systems.

These projects show that learning systems can be supplemented with artificial intelligence to automatically provide translations (e.g. Google Translate), analyze speech (e.g. automatic YouTube captions [9], Watson [10]), and annotate foreign language text (e.g. grammar analyzers, topic extraction). While each of these technologies provides some benefit for learners, these technologies are imperfect, especially in non-English languages where data is more sparse, and errors in output (e.g. errors in translations) might be detrimental to learners. While artificial intelligence may be problematic on its own, where there are errors, we can use

[8] fluentu.com
[9] youtube.com
[10] ibm.com/watson/services/speech-to-text

techniques from crowdsourcing to resolve those issues.

The idea that learners can be used to crowdsource learning content has been shown to be promising approach in building new learning support systems. Work by Kim et al. has shown learners can be prompted for information that can be used for improving learning conditions for other learners in how-to videos [57]. In their study, learners were prompted to generate summaries of segments of learning videos for the generation of step-by-step annotations. The website Duolingo[11] originally used language learners to translate foreign language material on the web while learning in the process [94], although the system has since been abandoned in favor of more engaging early language learning exercises. These systems show that learners can contribute to the development of learning resources while learning at the same time.

Considering these challenges and opportunities, in the following three chapters I discuss how we can design to reimagine how we can use a learner-centered approach with authentic materials to develop learning content (Chapter 3), design learning experiences (Chapter 4) and assess learning progress (Chapter 5).

[11] duolingo.com

CHAPTER 3

CREATING LEARNING SCAFFOLDS: LEARNER SOURCED CAPTION GENERATION

This section was written in collaboration with Solace Shen, Erik Andersen and Malte Jung. The work was published in the Proceedings of the Conference on Computer-Supported Cooperative Work and Social Computing, 2017 [24]. The introduction has been rewritten to better describe the work's relevance to this thesis, and a description of followup work that focused on applying the methodology to a new context has been added.

Native speaker materials are a promising resource for communicative language learning approaches, however these resources can often be difficult to use without support. When using these materials, learners will often encounter many new words and patterns, and the speech will likely be hard to parse. For example, consider foreign language entertainment videos. Many of these videos cover difficult topics such as historical events or technology, and the speech will likely seem very fast to the learner. Thus, by themselves, videos are difficult to use as learning resources, but with support learners can make sense of native speaker videos and learn effectively with them. Studies have shown that one of the most effective ways to learn through foreign language video is by watching the video accompanied by foreign language captions. This has been shown to be more effective than watching without captions or with translated subtitles [72, 64]. Although these captions are important to help learners get the most out of video resources, they are unavailable for many videos.

Traditionally, if a learner wanted to learn from a video without foreign language captions available, they would need to get those captions from an expert caption

creator (either a native speaker or longtime learner). Furthermore, existing caption creation systems are tedious to use and require extensive training to be effective [95]. This means that language learning communities are forced to rely on a select few experts for caption generation. To overcome these restrictions, we ask, what if learner communities could build their own captions while learning from and enjoying the video in the process?

In this chapter, I discuss a system where learners are given machine generated captions and caption editing tools. When learners notice mistakes in a caption, they can edit the caption on the video. We conducted a lab study to evaluate learning and engagement, and compared editing methods. The results suggest that there were similar amounts of language learning across conditions, despite participants making many corrections to the captions in the imperfect captions condition. This finding suggests that a caption correction task does not impair learning, so learners can improve their language skills while helping to build video caption learning resources. Furthermore, even novice learners were able to improve the caption quality, suggesting that even early learners can engage with imperfect resources and learn from those resources. This opens up a broader space of possible applications that allow learners to improve their language skills while helping build shareable learning resources.

6

The following is taken from Have your Cake and Eat it Too: Foreign Language Learning with a Crowdsourced Video Captioning System. /citeculbertson2017have.

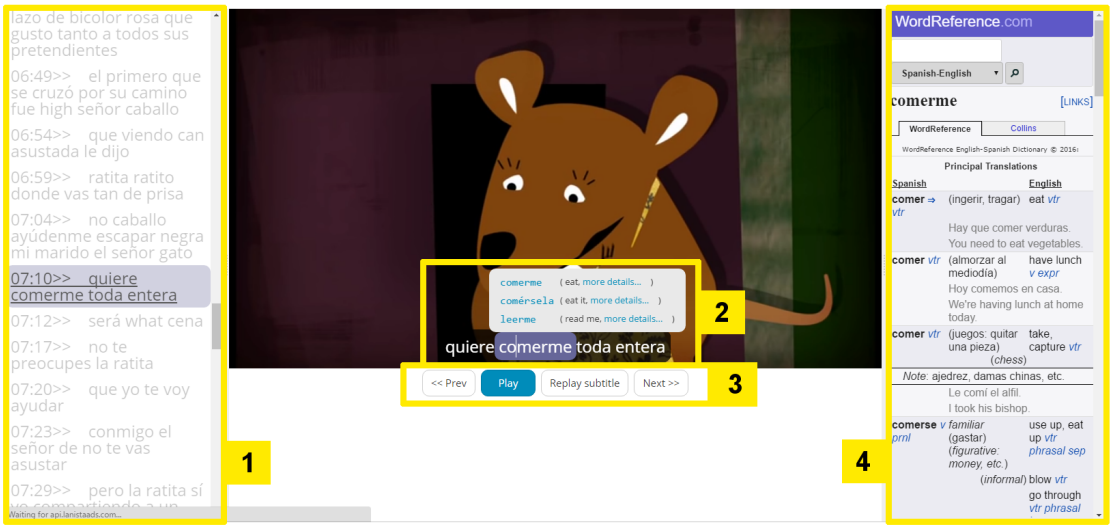

Figure 3.1: The caption video viewing and caption editing interface. Users can navigate through the video using the transcript (1), look at word definitions and edit the caption directly on the video (2), toggle video playback and jump one caption forward or back (3), and view expanded dictionary information using an external dictionary website (4).

3.1 Interface Design

The high-level goal of our system is to enable learning while participating in the captioning process. However, captioning is a difficult and time-consuming task. For example, for a learner beginning a subtitling task from scratch, "a single 5-min excerpt may take many hours for a novice to complete" [95]. Novice learners are handicapped by having incomplete or no knowledge of the caption language. Thus the key design focus is on minimizing the effort required to edit captions and providing enough scaffolding to support novice learners.

The system was developed as a website to maximize accessibility. A screenshot of the system along with descriptions of each element are shown in Figure 3.1. On the site, learners select a video and captions are generated using the IBM

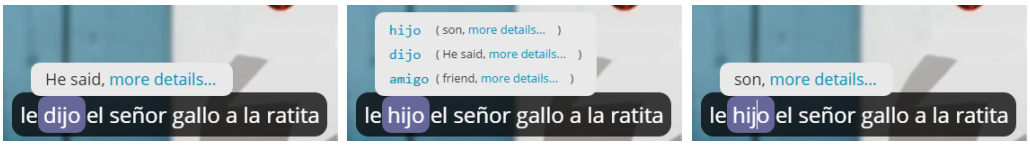

Figure 3.2: The 3 conditions used: (A) accurate captions, (B) imperfect captions with suggested alternative words correction task, and (C) imperfect captions with free response correction task. When participants clicked words in the subtitle, the word was highlighted and suggestion or translation information was displayed. In condition B, participants clicked the alternative Spanish word from the popup box to edit the caption. In condition C, participants typed words directly into the black area where the caption is displayed. In all cases, the displayed translation was generated with the Google translate API, and the 'more details...' link searched the selected word in the external dictionary.

Watson speech-to-text system[1]. The video is shown in the center with the captions overlaid on top of the bottom-center of the video. To edit the caption, learners simply click anywhere in the caption area to begin typing and changing words. If the speech-to-text system produces multiple possible words for an audio segment, these words will all be shown above the word containing the cursor (for example "comerme," "comersela," and "leerme" in area 2 of Figure 1). The word selected by the cursor will also show a translation generated by the Google Translate API, and the word can be searched in a dictionary by clicking a link "more details" next to the translation. Additional dictionary information is then displayed to the right of the video (area 5). If alternative words are displayed, each word will have a translation and dictionary link. Buttons (area 3) are displayed to allow users to move one caption forward or back and a transcript on the side (area 1) can be used to jump to a specific caption.

Producing words from scratch can be especially challenging for learners who may be unfamiliar with typing in that language. By beginning with the a machine

[1]http://www.ibm.com/

generated transcript, we greatly reduce the number of words that a learner would need to produce from scratch. We further reduce this by providing machine generated alternative words. Instead of typing the correct word, learner can instead just choose the correct word from a list.

Furthermore, word-by-word translations along with the video context enable learners to piece together the meaning of the caption. Work by Kovacs and Williams have shown that providing English translations for individual words can be helpful to learners to establish overall meaning [60]. However, because words can have slightly different meanings or nuances depending on the context, the translation will not always be entirely accurate. Therefore, we provide quick access to a full dictionary so that learners can view multiple possible translations and disambiguate the meaning.

3.2 Lab study: Assessing learning with different captioning workflows

In the design of this system, it was important to us to that learning occurred during the captioning process and that the task was engaging for learners. Otherwise, learners would have little incentive to engage with the system. Thus we conducted a lab study to evaluate the system. We had three key questions. First, (RQ1) are there significant differences in learning outcomes between using accurate vs. imperfect captions? If learning is significantly impaired by imperfect captions, then the system will not be a useful tool for language learners, and bringing captions from an external source may be more effective. Second, (RQ2) does the quality of the learning experience significantly differ between learning with accurate

vs. imperfect captions? Because the system will be learner driven, it's important that quality of experience is not significantly impacted by imperfect subtitles or learners will drop out of the system. And finally, (RQ3) does the complexity of the correction interface affect learning outcomes and quality of learning experience? If complexity of the editing task significantly influences learning or quality of experience, we can pursue designs that simplify the editing task.

To establish our expectation for RQ1 (are there significant differences in learning outcomes between using accurate vs. imperfect captions?), we consider the difference between active and passive learning. In the accurate caption condition, learners only need to understand the video, so we consider it a passive learning task. On the other hand, learners in the imperfect caption condition need to edit the caption to resolve disparities between what is being said and the caption, so we consider this an active learning task. Work by Michel et al.[70] has shown that active learning in the classroom is more effective than passive learning. Furthermore, specifically in the context of language learning, Gu and Johnson [43] found that Chinese learners of English were more successful in learning vocabulary when using active learning strategies. Therefore we expected that there would be more learning in the imperfect caption condition than the accurate caption condition.

To establish our expectation for RQ2 (does the quality of the learning experience significantly differ between learning with accurate vs. imperfect captions?), we looked at the work of Perez et al. [78] which looked at the use of keyword-only captions (where only keywords were displayed) in comparison with complete captions for foreign language learning. They found that the keyword-only captions and complete captions result in similar learning gains, but participants reacted negatively to the keyword-only captions. Therefore, we expected that while participants

would learn from watching video with imperfect captions, they will perceive the quality of their video learning experience as worse when presented with imperfect subtitles.

To establish our expectation for RQ3 (does the complexity of the correction interface affect learning outcomes and quality of learning experience?), we looked at work by Cades et al. [13] which has shown that the interruption task complexity affects primary task performance. Therefore, we expected that a less complex caption editing method (e.g., one in which selectable options are offered) will result in better performance than a more complex editing method (free open-ended response).

3.3 Experiment design To assess the effects of different captioning systems, this study used a between-subjects design with 3 conditions. As shown in Figure 3.2, the three conditions used were: (A: accurate) accurate captions, (B: suggested-alternative) imperfect captions with suggested-alternative-word editing, and (C: free-response) imperfect captions with free-response editing. In all conditions, participants viewed a Spanish language video entitled La Ratita Presumida and the dictionary link searched the word in SpanishDict[2]. The same video was used in all conditions to control for differences in interest and difficulty that may come with different videos. In the accurate caption condition, participants were presented with error-free Spanish caption for the video. This served as the baseline condition. In the suggested-alternative imperfect caption condition, participants were presented with caption that contained errors and offered a list of phonetically similar alternatives they could choose from to correct the caption. Word options for this condition were generated by the IBM Watson speech-to-text system[3]. In

[2]http://www.spanishdict.com/
[3]http://www.ibm.com/

{ una, a, al, ha }
{ errática, ratita, ratito, regatista, raquítica }
{ quiero, quiere, quédate }
{ pero, pino, pedro, pido, peer }

Figure 3.3: Sample errors generated by speech-to-text system

the free-response imperfect caption condition, participants were presented with the same imperfect caption, but instead of having selectable options, they had to type words to correct the caption. For both the suggested-alternative and free-response imperfect caption conditions, the imperfect caption was generated by beginning with a correct transcript and introducing a 19% word error rate, where the incorrect words were drawn from the machine generated word alternatives.

```
foreach word ∈ correcttranscript do
    rand = random number between 0 and 1;
    if rand <0.2 then
        add random incorrect word from ASR to bad transcript;
    else
        add word to bad transcript;
    end
end
```

A maximum of four alternatives were displayed, but as few as one alternative would be displayed if the speech-to-text system did not identify any alternative possibilities for a given word. Sample errors are shown in Figure 3.3.

In all conditions, participants were told that the captions were machine generated and may contain errors. We did not tell participants the purpose of the subtitle correction task because we wanted to control for framing effects in measuring engagement and understanding intrinsic motivation.

3.2.1 Experiment procedure

We used the following procedure. After giving assent, participants were asked to sit in front of a laptop with headphones on. Participants completed a brief survey about their demographic information and Spanish learning background, along with a Spanish vocabulary and reading speed test. Then an experimenter explained to participants how to use the caption editing interface. The experimenter told participants that they were free to edit the captions if they wished and could take as long as needed to watch the entire 11 minute video. While watching, participants could click the Spanish words in the captions to see their English translations and could edit mistakes in the captions. After completing the video, participants were automatically redirected to a post-test on vocabulary, reading speed, and comprehension. Finally, participants completed a survey about their experience.

3.2.2 Participants

Participants were recruited using a university recruitment system and email lists. Native speakers of Spanish were not permitted to participate. A total of 54 participants (55% female) were recruited and 49 participants were included in the final analysis. The 5 participants that were not included experienced technical difficulties which invalidated their data. The age of participants ranged from 18 to 39. Participants were randomly assigned to conditions (accurate [A] = 18, suggested-alternative [B] = 16, free-response [C] = 15).

Pearson Correlations

	Self-rated Spanish skill	Vocabulary (pre)	Comprehension	Reading speed ratio change	Effectiveness perception	Frustration	Video interesting and engaging	Vocabulary (change)	Edits
Self-rated Spanish skill	—	0.721***	0.482***	-0.360*	0.460***	-0.215	0.076	0.394**	0.174
Vocabulary (pre)		—	0.377**	-0.338*	0.408**	-0.088	0.167	0.433**	0.111
Comprehension			—	0.031	0.250	-0.259	0.276	0.473***	0.299*
Reading speed ratio change				—	-0.259	0.249	-0.116	0.053	0.079
Effectiveness perception					—	-0.323*	0.370**	0.263	0.134
Frustration						—	-0.253	-0.338*	-0.009
Video interesting and engaging							—	0.137	0.060
Vocabulary (change)								—	0.121
Edits									—

* p < .05, ** p < .01, *** p < .001

Figure 3.4: Table of correlations for measures of interest.

3.3 Results

3.3.1 Measures

Learning measures

Learning was measured using a vocabulary translation test and a reading speed test.

In the vocabulary test, 20 Spanish words were picked randomly from the video and learners were asked to type the English translation for each word. The same 20 words were used on the pretest and posttest, but the order was randomized. A pretest and posttest score were calculated by counting the number of words correctly translated. Number of words learned was calculated by subtracting the pretest vocabulary score from the posttest vocabulary score.

For the reading speed test, a standard sentence processing test was used [35].

Words appeared one word at a time, and in the end the learner was asked whether the sentence they just read was correct or not. This is a useful measure for testing language ability, because, like listening to speech, learners must process words as they read them or they will be unable to understand the entire sentence. We calculated reading speed on 30 sentences (15 from the video and 15 not in the video) before and after the video captioning task. Sentences not from the video were included because, unlike vocabulary learning, we expect that grammar learning from the video should generalize beyond just the sentences from the video. Because individuals have differences in general reading speed, we used the percent change in reading speed as our reading speed learning measure.

Quality of experience measures

Four measures were used to gauge the learner's quality of experience. For the first three measures on participants' perceived quality of experience, participants rated on seven-point scales (describes the task... "not well at" [1] to "extremely well" [7]) how effective for language learning they found the task to be, how engaging and interesting they found the video to be, and how frustrating they found the task to be. Finally, we measured comprehension using 20 true/false/unsure questions about the details of the video story. The comprehension score was calculated as the number of incorrect responses subtracted from the number of correct responses, and "unsure" answers were ignored. We used this scoring system because it gives us a more precise measure by discouraging guessing in cases where the learner is uncertain.

Editing behavior

In order to measure caption editing behavior, we measured the number of times the learner changed a word. This change could either be through clicking on a menu interface (suggested-alternative condition) or typing (free-response condition). Furthermore, we measured whether this change was correct or not.

Final caption accuracy measures

Two final sets of captions were generated by aggregating the changes made by participant groups in the suggested alternative condition (B) and the free response condition (C). In each condition, the final transcript was generated by beginning with the incorrect caption set. Then, for each word, if no participant made a change to that word, the word remained the same as the one in original incorrect caption set. If at least one participant made a change, the new word was determined by taking the change made by the majority of participants. In the case of a tie, a word from the tied-majority words was chosen pseudo-randomly. Finally, a word error rate (WER) was calculated for each final transcript. The average improvement made by each participant was also measured. Note that this is not equal to the final WER divided by the number of participants because many participants made the same changes.

Furthermore, in order to better understand the viability of the system for novice learners, we looked at the final accuracy of captions generated by just accounting for changes made by learners who reported little to no Spanish experience (1 on a 7-point scale).

3.3.2 Findings

A table of results are presented in Figure 3.6 and learning outcome graphs are shown in Figure 3.9. Participants in all conditions showed significant ($p < .001$) improvements in reading speed between the pre- and post-test. The mean percent change was -26.9% for the accurate condition, A, (SD = .30), -27.4% for the suggested-alternative condition, B, (SD = .19), and -19.6% for the free-response condition, C (SD = .25). A one-way ANOVA indicated that the differences between conditions were not statistically significant for percent change in reading speed, $F(2, 45) = 0.44$, $p = .650$, partial $\eta^2 = .019$. Participants in all conditions also showed significant ($p < .001$) improvements in vocabulary scores between the pre- and post-test. Thus we replicated earlier findings that video learning leads to learning gains. Participants on average increased scores by 3.44 points in the accurate condition, A, (SD = 2.48), 1.88 points in the suggested-alternative condition, B, (SD = 1.93), and 2.40 points in the free-response condition, C (SD = 2.29). A one-way ANOVA indicated that the differences between conditions were again not statistically significant for change in vocabulary score, $F(2, 46) = 2.15$, $p = .128$, partial $\eta^2 = .085$. These results suggest that as expected, learning, in terms of reading speed and vocabulary improvements, was not significantly different with either accurate or imperfect captions. In the suggested-alternative condition, 17% of words learned were learned after the participant edited that word and in the free-response condition, 15% of words were learned after the participant edited that word. Making a correction did not increase the probability of learning a word.

Participants also largely reported similar quality of the learning experience across conditions (see Figure 3.6). One-way ANOVAs indicated that the differences between conditions were not statistically significant for how effective participants

word	Acc.	Sug.-alt.	Free-res.	Correction
ayúdeme	22%	18%	20%	no
barrio	33%	12%	13%	yes
caballo	11%	18%	13%	no
cada	11%	6%	26%	yes
casarte	16%	6%	26%	yes
cerca	5%	0%	6%	yes
corriendo	16%	6%	20%	yes
cruzó	0%	6%	0%	no
encontró	11%	6%	6%	no
entera	38%	12%	20%	no
lazo	55%	37%	20%	yes
llegaron	0%	0%	6%	no
marido	22%	25%	33%	no
paseando	0%	0%	13%	yes
pendientes	0%	0%	0%	yes
presumida	66%	31%	13%	no
prisa	11%	18%	20%	yes
sacó	0%	6%	6%	yes
uñas	27%	6%	0%	yes
viendo	11%	6%	0%	no

Figure 3.5: Percent of participants that learned each of the words on the vocabulary test.

found the task to be for learning, how engaging/interesting they found the video to be, and how frustrating they found the task to be. However, a statistically significant difference between conditions was found for video comprehension, $F(2, 44) = 4.98$, $p = .011$, partial $\eta^2 = .185$. Post hoc comparisons using the Tukey HSD test revealed that the mean video comprehension score was significantly higher with accurate caption in Condition A ($M = 10.72$, $SD = 6.01$) than with imperfect caption in Condition B ($M = 4.81$, $SD = 5.90$), $p = .011$. Descriptively, the mean video comprehension score was also higher with accurate caption in Condition A than with imperfect caption in Condition C ($M = 6.39$, $SD = 4.75$); this difference however was not statistically significant, $p = .100$. Quality of experience graphs are shown in Figure 3.10 These results suggest that contrary to expectation, partic-

ipants did not perceive the quality of their learning experience to be significantly worse when presented with imperfect captions, although the imperfect captions seemed to have negatively affected their comprehension of the video.

No statistically significant differences were found for any of the learning or quality of experience measures between Condition B and C in the above ANOVAs and if applicable, post hoc comparisons. This suggests that the complexity of the caption editing method, whether with selectable options or requiring open-ended responses, did not seem to affect participants learning outcomes or experience in this study.

Results of the accuracy analysis are shown in Figure 3.7 for all learners and Figure 3.8 for novice learners. Aggregating learners' corrections overall improved final caption accuracy by 13.5% (from 19% to 5.5% WER) in the free response condition, and an even larger improvement of 17.2% was observed in the suggested-alternative condition (from 19% to 1.8% WER). While no improvement was observed for novice learners in the free response condition, a 10.1% improvement was observed in the suggested-alternative condition (from 19% to 8.9% WER). These results suggest that our caption correction system was effective in producing accurate captions from aggregated learner inputs, and that novice learners were able to benefit more from the scaffolding feature of suggested alternative words.

To gain insights for system improvements, we conducted correlation analyses to explore potentially important relationships (see Figure 3.4). One relationship of interest that emerged is a moderate correlation between measures of preexisting language ability and how effective participants found the the interface for learning ($r = 0.460$), as well as vocabulary learning ($r = 0.394$) and comprehension ($r = 0.482$). This indicates that, according to our measures, the learning task was

	Accurate		Suggested-alternative		Free-response			
	M	SD	M	SD	M	SD	F	p
Vocabulary change	3.44	2.48	1.88	1.93	2.40	2.23	2.15	.128
Reading speed ratio change	-0.27	0.30	-0.27	0.19	-0.20	0.25	0.44	.650
Effectiveness perception	4.33	1.20	3.77	1.46	3.75	1.42	1.00	.376
Video interesting and engaging	4.56	1.29	3.84	1.14	4.03	1.06	1.69	.197
Frustration	2.89	1.23	3.94	1.77	3.36	1.34	2.19	.089
Comprehension	10.72	6.01	4.81	5.90	6.39	4.75	4.98	.011
Difficulty perception	4.25	1.62	4.03	1.38	3.62	1.34	0.71	.497
Time on video	15.70	7.83	18.32	10.15	20.00	9.77	0.91	.409
Edits	-	-	53.31	58.12	42.13	39.91	1.18	.316

Figure 3.6: Mean Ratings/Score (Standard Deviations) of Quality of Experience Measures

Condition	N	Final WER	Avg. WER Improvement	Correct edits	Incorrect edits
Suggested-alternative	16	1.8%	4.1%	634	50
Free-response	15	5.5%	1.6%	241	31

Figure 3.7: Final caption accuracy for all learners

Condition	N	Final WER	Avg. WER Improvement	Correct edits	Incorrect edits
Suggested-alternative	8	8.9%	3.3%	251	23
Free-response	7	19%	0%	0	5

Figure 3.8: Final caption accuracy for novice learners

generally more effective for learners with some experience.

3.4 Discussion

3.4.1 Research questions

No significant differences in learning outcomes were found using accurate vs. imperfect captions (RQ1)

There were no significant differences between learning outcomes across conditions, but all conditions showed evidence of learning. Therefore, we conclude that while the caption editing task did not improve learning, it also did not impair it. Although this does not match our expectation, this finding aligns with the findings of Semke [83] that shows corrections to writing do not influence writing improvement, but rather the act of writing leads to improvement. Semke [83] found that having students correct their own foreign language writing rather than having teachers

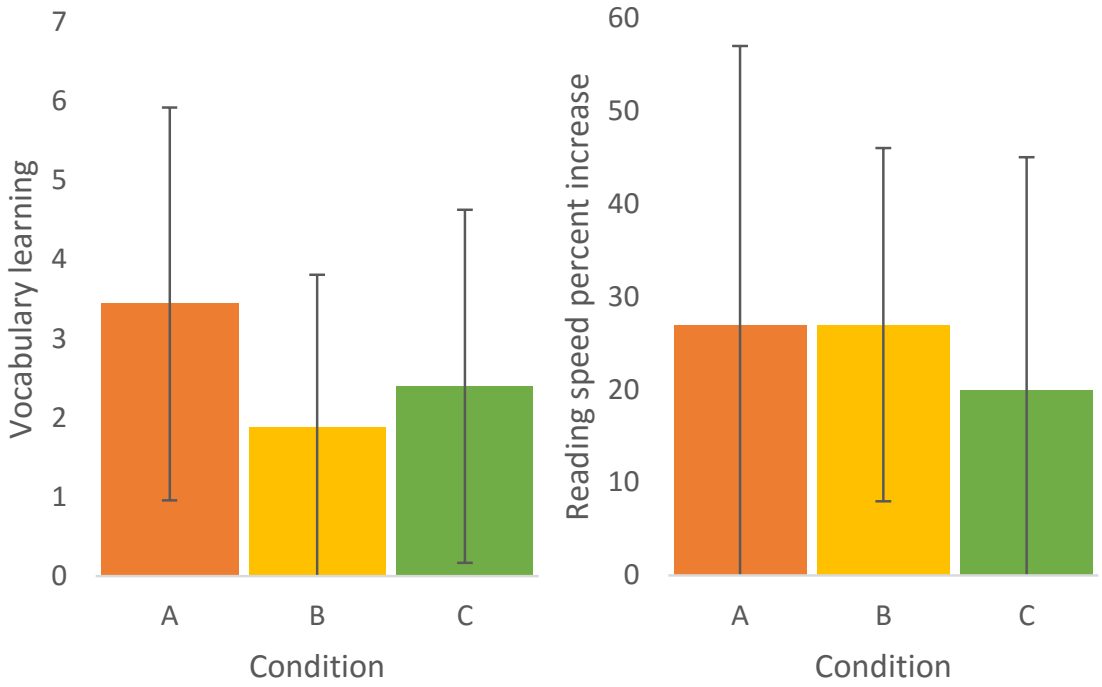

Figure 3.9: Learning measures are shown above for the three conditions (A: accurate captions, B: suggested-alternative imperfect captions, C: free-response imperfect captions). We found no significant difference between learning measures across conditions. Each error bar indicates +- 1 standard error.

explicitly marking errors made no difference in writing improvement over a 10 week period. Similarly, learning by our participants was likely driven by the act of watching the foreign language video rather than the act of correction.

It is also possible that learners in all conditions were actively learning. Although we initially expected that learners in the imperfect caption conditions would be learning more actively, it likely that although learners in the accurate caption conditions did not need to make edits, learners still needed to work actively to understand the story. This finding suggests that in distinguishing between active and passive learning tasks, the structure of the task is less important than the mental processes of the learner. Because the study took place in the lab, we expect that all of the learners would feel obligated to engage with the task, so

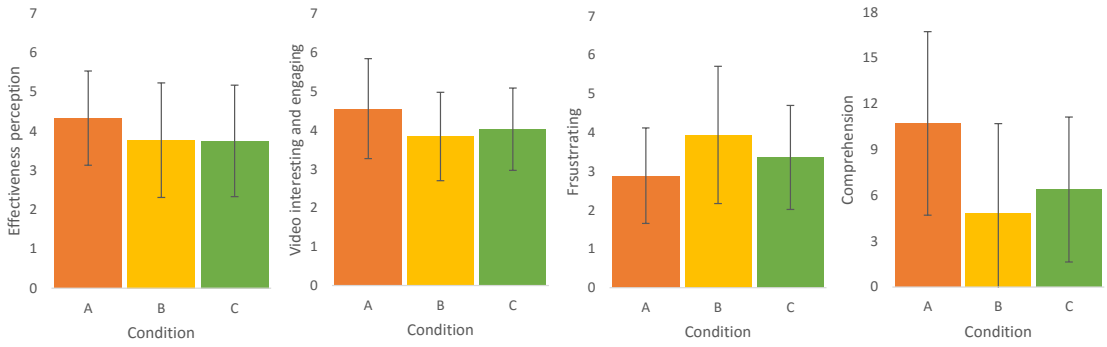

Figure 3.10: Quality of experience measures are shown above for the three conditions (A: accurate captions, B: suggested-alternative imperfect captions, C: free-response imperfect captions) with each error bar indicating +- 1 standard error. Significant differences were only found in the comprehension score. Note that the standard errors are quite large. In future work, it would be beneficial to repeat the study with more participants to reduce the errors.

it is possible that we would see differences in learning activity in less controlled conditions.

Despite having no significant difference across conditions, learning did occur in all conditions. Therefore, we believe that our system can enable learners gain language skill while contributing to the generation of accurate video captions.

The quality of the learning experience did in some ways significantly differ between learning with accurate vs. imperfect captions (RQ2)

We found that comprehension was lower in the suggested-alternative (B) condition and free-response (C) condition than accurate caption (A) condition, but the difference was only significant between the accurate caption (A) condition and suggested-alternative (B) condition. No significant differences were found across conditions for other quality of experience measures. This means that although

learners with accurate captions understood more, they felt similarly engaged and frustrated. We suggest that the loss of comprehension with imperfect captions is due to the increased focus on low-level details in the editing conditions. Learners needed to focus on individual words rather than the high level story in order to complete the task.

In future work, it will be important to explore whether learners feel a loss of understanding is detrimental to the usefulness of the system. If the primary goal of the learner is to gain language knowledge, this may be acceptable, but in scenarios where complete comprehension is important, the system may not be appropriate. It may also be useful in the long-term to alternate between focusing on high-level comprehension and low-level attention to detail.

We also looked at comments written by frustrated participants describing why they felt frustrated. Because we only gave participants a limited picture of the overall system, many participants were confused by the presence of the errors or felt they could have learned more with accurate captions. For example, one participant wrote I didn't really understand the whole "some subtitles aren't correct" thing. In future work, we plan to explore ways to make the rationale and benefits of the system more clear to learners. However, it should be noted that while some participants indicated frustration with the editing task, we did not measure any statistically significant overall differences in frustration between correct and incorrect caption conditions.

The complexity of the correction interface did not affect learning outcomes and quality of learning experience (RQ3)

We found no significant difference between the two interface setups that we tested. More edits were made in the C condition (open ended editing) than the B condition (suggested alternative editing), but the difference was not significant. Originally we expected that the suggestion interface would offer a simpler and less taxing way to edit the caption, but our study suggests that there is no difference between this and open ended editing. We believe that this may have been because additional time needed to be spent reading the alternative words and their definitions which turned out to be roughly equivalent to the amount of time that was needed to type the word.

3.4.2 Beginner and intermediate learners

Our exploratory correlation analysis revealed that the learning task was generally more effective for learners with some experience. However, work in language acquisition has shown us that learners need to learn to parse and chunk sounds in a language before they can learn higher order processing such as vocabulary learning [20]. Although we did not measure this, learners were listening and trying to understand the foreign language audio, so naturally they would gain some skill in sound processing through this task. Furthermore, despite not remembering all of the vocabulary, beginning learners reported similar interest and engagement with watching the video. Although future work will need to tease out the nuances of video learning and skill level, we feel that our findings indicate that with adequate scaffolding video learning can be productive and engaging for learners of any level.

Our accuracy analysis showed that novice learners made correct edits in the suggested-alternative condition, but participants were unable or unwilling to make correct edits in the free-response condition. We believe that the additional scaffolding provided by the suggested-alternatives gave novice learners enough confidence to make changes to the captions.

3.4.3 Motivations for caption editing

In order to guide future design, we explored what motivated learners to edit captions using comments from our survey. In the survey we asked participants: "what motivated you to edit the subtitles?". While some participants indicated that the research study was the primary motivator, others noted that the errors in the captions prevented them from understanding the story. In order to better understand the video, they corrected the caption and read through it again. Others made edits because they were bothered by the mismatch of audio and text, or the story context and the caption meaning. This suggests that these participants found satisfaction in having a correct end-product. Both of these orientations suggest different design approaches. For example, if the primary motivator is understanding the story, additional scaffolding could be provided to learners to help them reach a complete understanding of the video. On the other hand if learners are motivated by perfecting the caption, parts of the caption that are likely to contain errors could be made more salient.

3.4.4 Corrections by learners

Our final caption accuracy results suggest that our caption correction system can produce accurate crowd-sourced captions, and that the scaffolding feature of suggested alternative words provided critical support for novice learners. Despite the fact that learners did not know the words and grammar beforehand and the errors were not highlighted in any way, even learners with no experience were able to make caption corrections in the suggested-alternative condition. The responses to the post-task survey indicated that participants used a combination of cues to establish the existence of an error and then to make a correction. Participants used word translations along with the audio context ("many words did not align with what the narrator was saying"), the visual context ("the sentence did not quite express what was happening in the visual scene"), and the narrative context ("get [the caption] back on track with the story"). These results indicate that with adequate scaffolding and the right video, even absolute novices can learn and contribute to caption generation.

3.5 Limitations

Although the system is intended for any video, it should be noted that the same video was used for all participants. We expect that the match between a learner's skill level and interests with a video will influence learning and quality of experience. The video that we did select was created for children and contained a lot of repetition. Some participants pointed out that this helped them with learning. While the use of just one video does not adequately reflect how we intend the system to be used in practice, our finding, that measurable learning gains occurred

despite the use of just one video, is providing us with additional confidence about the effectiveness of the system for language learning. Future work should more carefully examine the effect of video difficulty and fit on learning outcomes.

Learning through caption correction was a novel method for all participants and, given the short duration of the study, learners may not have totally acclimated themselves to this learning method. Although participants did not indicate difficulty learning to use the interface, learning effectively through correction may require a shift in learning strategy. Future work should examine whether the system is effective over longer periods of time, and how utility perception and engagement change with time.

Learners may have been primed to learn the words from the pretest. However, this would not affect differences between conditions as all participants were primed in exactly the same way. Furthermore, this type of priming would not be unusual in real use scenarios. Often teachers prepare students with vocabulary lists before introducing new dialogues, and we could easily implement similar learning methods with our system by asking learners to pay special attention to words of interest in the video. We have added discussion for this to the limitations section.

Furthermore, the insignificant learning differences across conditions may have been in part due to our small sample size. The power of our study was too low to detect potential small differences. However, it should be noted that learning was measured to be significant (pre to post) in all conditions (though for reading speed specifically we would expect increased performance in repeating the test). Future work should further explore the difference between learning while editing captions and learning through other video engagement methods with a large sample.

Finally, measuring language learning is incredibly challenging given the complex, intertwined processes that are necessary for language comprehension. Williams and Thorne identified abilities to "listen attentively, recognize and fully absorb the content" [95] were essential for students to effectively produce subtitles. Our measures of vocabulary, reading speed and comprehension provide only limited windows into the the learning processes that are taking place. Future work should investigate other ways to evaluate language learning through captioning.

3.6 Further work

Following the lab study, a fully automated system was developed. Using the system, learners can upload a video or choose a YouTube video. The audio is extracted and processed to generate the initial set of captions. Multiple users can then simultaneously edit a set of captions. Although timestamps are generated automatically for timing captions, these can be adjusted. A visualization of the audio from the video is displayed to assist with caption time adjustments.

Furthermore in related work, we showed that the paradigm of learner modification based on a machine generated starting point can be used in other contexts as well. In this work we extracted the text from a popular Japanese game (Final Fantasy XIV). The game contains text in both English and Japanese and players can switch the language in the game as they like. Because of this and the ability to interact with real Japanese speakers online, some players have user the game to learn Japanese. However, learning with the game can often be challenging because Japanese uses Chinese characters or kanji which can have many phonetic readings that change based on context. Some tools exist to label kanji with appropriate

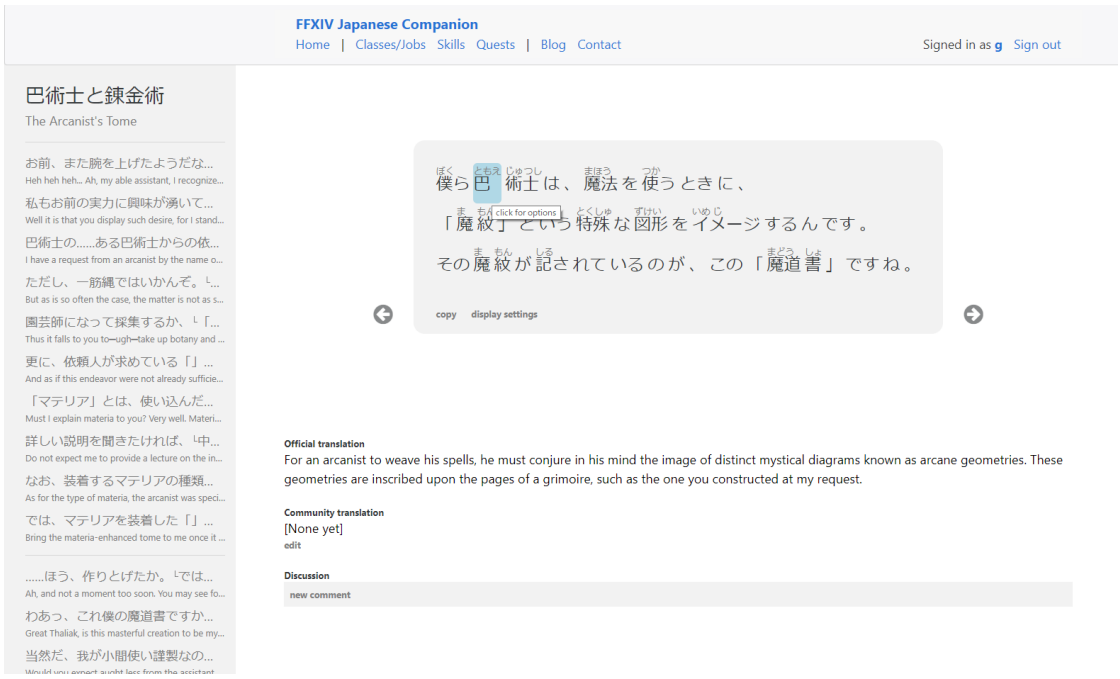

Figure 3.11: Learners can view text from the game with phonetic annotations and correct those annotations which have errors.

readings, but these tools are imperfect and are especially prone to error in fantasy contexts where uncommon vocabulary is used.

To address this challenge, a website was developed where the game text was posted, and users could contribute correct readings for kanji of the game text. This information was integrated with a tool which shows players the correct readings for kanji in the game as they play. Again we show that by using a starting point that is mostly right, we can enable learners to contribute to learning resources while learning in the process.

3.7 Conclusion

In this work, we presented a system to enable foreign language learners to learn while correcting video captions that could then be used by a wider learning community. Our findings suggest that although learners' comprehension was reduced by the editing task, it did not influence learning or engagement with the video. Comments from our usability indicated that contributing to building captions was motivating for some learners. Given this combination of findings, we envision a system where learners motivated by social contribution and learners motivated primarily by their own learning could both learn and contribute by collaborating on generating captions. Where previously learners would need to rely on external sources to generate captions for learning videos, we have shown it is possible, with the help of artificial intelligence, for learners to build their own learning content.

Although learning through the captioning process is an important first step, there are other important language skills for learners to practice. Using materials generated through the approach described in this chapter, we have new design opportunities for helping learners practice these important language skills. In the next chapter, I will describe a system for practicing an essential language skill, speaking, and show how captions (like the ones generated from the captioning system) can be used as a key component in this design.

CHAPTER 4

DESIGNING EFFECTIVE LEARNING INTERACTIONS WITH NATIVE-SPEAKER MATERIALS

This section was written in collaboration with Solace Shen, Malte Jung and Erik Andersen. The work was published in the Proceedings of the Conference on Human Factors in Computing Systems, 2017 [25]. The introduction has been rewritten to better describe the work's relevance to this thesis, and a description of followup work that focused on design for classroom use has been added.

Even with annotated resources such as captioned videos, we need to design learning experiences around these materials that help to focus learners on important language learning skills. Speaking skills, especially understanding and producing nuanced speech (i.e. pragmatics), are fundamental to foreign language proficiency. Although videos are an excellent source of pragmatically rich content, only limited pragmatic competence can be gained from passive video watching [93]. Active engagement with the videos that involve actual practice would lead to more optimal acquisition of pragmatic competence.

A simple, yet effective, form of active engagement with foreign language videos is repetition. Studies have shown that merely repeating a sentence requires a learner to be able to completely process a language [28]. Existing language learning tools like DuoLingo[1] and Rosetta Stone[2] do not prioritize speaking and existing video learning tools (e.g. [60]) use text rather than speech. Creating tools that focus on repetition of language, and integration of repetition into workflows involving native speaker materials could open new possibilities for language learning

[1] https://www.duolingo.com/
[2] www.rosettastone.com

tools.

In this chapter, I discuss Seiyuu-Seiyuu, an online video-based learning tool that takes a step towards these goals. Seiyuu means voice actor in Japanese. In Seiyuu-Seiyuu, users suggest videos to watch through a crowdsourced website by linking to YouTube videos. Then, Seiyuu-Seiyuu allows users to repeat utterances they hear in the video, and what's more, to take on the role of an voice actor and speak with paralinguistic cues such as intonation, pitch, etc. Seiyuu-Seiyuu takes advantage of Google's speech recognition technology to recognize what the user is saying. When using videos for which a transcript has been uploaded, the system allows users to see how much of the video they have correctly repeated.

In an online evaluation study of 27 participants, we compared this system to a text-based translation interface similar to what many participants use in their language learning already. In the study, learners used both the voice and text interface. We found that learners searched 53% more words using the voice interface than the text interface. Furthermore, learners who used voice first conducted more than twice as many total searches with voice and text, indicating an ordering effect from using the voice interface. Furthermore, our qualitative findings support previous research that shows the value of learning with foreign language videos, and suggests that the voice interface is better suited for learning practical conversational skills.

The following is taken from Facilitating development of pragmatic competence through a voice-driven video learning interface. [25]

Figure 4.1: The interface with game features provides feedback when learners sayid phrases correctly (1) over transcribed and translated text (2). A progress bar and text displays how much of the video the learner has correctly repeated (3). Learners can add utterances to their library or remove them using buttons (4), or upload transcripts and adjust how text is displayed through the settings (5). When available, a transcript is also displayed to show how much of the video a learner has repeated, and help learners find new words and phrases to listen to (6). Screenshot taken from Ode to Joy on YouTube (https://www.youtube.com/watch?v=4wGpu56WQGQ).

4.1 Design

The most common existing methods for learning a foreign language make use of carefully designed learning materials, which, we argue, fail to provide the breadth of experiences that are necessary to gain situational fluency. Use of authentic materials has been shown to enable pragmatic competence learning. However, using materials designed for native speakers can be incredibly challenging because there are many unfamiliar words and structures in these materials and finding out the meaning of these words and structures is difficult. Therefore, our primary design

question was: how can we design to make authentic foreign language materials more accessible?

We observed three key opportunities which inspired the design: (1) the internet is home to countless free authentic foreign language videos, (2) repeating phrases or dialogues from videos is a natural activity and helps improve oral proficiency, and (3) speaking bypasses the need to type which is often very challenging and time consuming in a foreign language.

(1) Freely available videos on YouTube include 76 languages and over 100 million videos[3]. Some countries also have their own streaming video sites (i.e. China - http://www.youku.com/) which can increase the number of videos even further.

(2) In adult learning, many educators use oral repetition as a central learning excercise (e.g. [91]). Furthermore, listening to foreign language before speaking (known as word priming) has been shown to improve recognition and pronunciation of those words [90].

(3) In our survey of authentic material use by foreign language learners, we found that of those that used materials designed for native speakers, 45% reported using a text based tool such as a dictionary or Google translate in order to learn from the material. Furthermore, many learners use videos with flashcard systems such as Anki[4] which requires transcribing text from videos.

Considering these opportunities, we developed an interface for watching any foreign language video while the learner can speak words or phrases to see them transcribed and translated below the video.

[3]https://www.youtube.com/yt/press/statistics.html
[4]anki.net

Learners use the system by first selecting a video to learn from. For example, a learner of English might choose the television program Friends. As the learner watches the program, they listen carefully for words and phrases they can pick out and then repeat them. For example, maybe the learner hears the phrase "Chandler, I sensed it was you." but is unsure of what "sensed" means. The learner would then hold the spacebar to pause the video and repeat the phrase aloud. The speech recognition system would recognize and display the text for the phrase. Below the transcribed text, a translation would appear in the learner's native language. After the learner reads the transcription and translation, they release the spacebar and continue watching the video.

The system was implemented using node.js as a backend. Speech recognition was realized using the Speech Recognition Webkit built into Google Chrome. Translation used the Google translate API.

4.1.1 Website

To improve long-term learning and engagement, the web version of the system includes additional features to support existing language learning practices and community building. To support long-term learning, spoken utterances can be saved to a history and edited. A plugin was also developed to allow learners to sync saved phrases with the spaced repetition system Anki[5].

To better engage learners with the system, game features were added. Learners can choose to upload a caption file, which allows the system to check for the accuracy of utterances. When captions are available, learners receive a score based

[5]anki.net

Figure 4.2: When visiting the website, learners choose a language (1) and can then view videos that other learners watched in that language (2). Learners can also choose their own video from YouTube or their computer (3). Links to Youtube become visible for all users, but personal videos are only visible to the user who uploaded them.

on how much of the video they are able to repeat and learners can watch the same segment multiple times to increase overall progress. Furthermore, a transcript of the video is shown on the side to indicate which words have already been spoken and which words the learner still needs to say. The system is shown in Figure 4.1.

The website includes a popular page where recently viewed YouTube videos are displayed and learners can view their progress as shown in Figure 4.2. Learners can also choose to add their own video from YouTube or their hard drives. If learners choose a new YouTube video, the video gets displayed on the popular page. Videos from users' hard drives videos are not uploaded to the server or displayed on the popular page.

4.2 Evaluation

4.2.1 Field study

The site was announced on the reddit LanguageLearning forum[6] as well as individual sub-reddits for Japanese[7] and Spanish[8]. Data was collected through usage logs. A total of 130 participants tried the system and 71 learners spoke 10 or more phrases. Learners that spoke 10 or more phrases spoke an average of 71 utterances. Users uploaded 22 new YouTube videos (in French, Spanish and Japanese) and used 6 unique media from their hard drives. This suggests that the tool can function in the wild.

Furthermore, since some videos may be more effective for pragmatic learning

[6]reddit.com/r/languagelearning
[7]reddit.com/r/LearnJapanese
[8]reddit.com/r/learnspanish

than others, we believe that identifying and sharing effective resources is an important task. Our findings about learner use of our system suggest that the tool could provide motivation for learner-sourced resource evaluation and sharing.

The results indicated that the system has potential to function as an independent learning tool, but we wanted to do a more systematic exploration in order to better understand how the tool compares to existing tools.

4.2.2 Formal Evaluation

To gain insight into the usability and effectiveness of our system compared to other video learning methods, we conducted an online study with foreign language learners. Originally, the study was available in six languages, but we only had participants use Spanish, French, and Chinese (Mandarin). In our survey of language learners, we found that learners most frequently used Google Translate to learn from native speaker materials. Therefore, to examine the effectiveness of Seiyuu-Seiyuu we developed an interface as a control that allowed learners to type into a textbox upon which translations appeared below using the Google translate API as shown in Figure 4.4. We used a within subjects design where approximately half of the participants first used the speech interface for video learning and the remaining participants first used the text interface. For all languages, realistic dramas or comedies were chosen for learners to watch. The video was different for each language. This is because culture is an essential element of pragmatic learning, so it was important to us to choose videos coming from cultures where each language was spoken. The videos used are shown in Figure 4.5.

Participants clicked on a link which redirected them to a webpage where they

Figure 4.3: In the voice interface used in our evaluation, learners were given instructions on how to use the system (1), and spoken phrases appeared below the video (2) with a translation below the utterance (3). In Japanese and Chinese, pronunciation was displayed beneath the characters. Since the speech recognition was not always accurate, the "more" button (4) could be clicked to show alternatives from the speech recognition system. Screenshot was taken from Keikon Dekinai Otoko on YouTube (https://www.youtube.com/watch?v=dX8vYhztrxM).

spoke a test phrase into their microphone in order to verify that speech recognition was working properly. Participants then chose a language and completed a short survey about their prior experience with the chosen language. Next, participants were randomly assigned to either use the speech interface first or the typing interface first. When using the speech interface, text was displayed to indicate that participants should hold the spacebar to pause the video and begin speaking. After pressing the spacebar, the interface would indicate that they should begin speaking

Figure 4.4: In the typing interface used in our evaluation, learners were given instructions below the video (1), could type word or phrases into a text field (2) and translations would appear below after the learner stopped typing (3). Screenshot taken with Sur Le Fil on YouTube (https://www.youtube.com/watch?v=bapP3JM3SZA&t=314s).

Language	video
Spanish	Mi Coraz Tuyo
French	Sur Le Fil
Chinese (Mandarin)	(huan'le'song), Ode to Joy

Figure 4.5: Learning sources for surveyed learners.

as shown in Figure 4.3. As the participants spoke, their utterances were recognized and translated below the video. In the typing condition, the interface indicated that participants could pause the video by clicking on it, and participants could type words and phrases to see their translations. In each case, the video segment was 10 minutes long. Following the first video segment, participants were asked

to rate difficulty, usefulness and enjoyment as well as recall words from the video with their surrounding contexts. Participants then watched a different 10 minute segment of the same video using the interface that they did not use in the first part of the study. Following the second task, participants were asked to report the same information as in the first part. Finally, participants were redirected to a survey where they provided demographic information and were asked to discuss their learning experience and compare the two interfaces.

Participants

Participants were recruited through a campus research system. Participants were compensated either $10 or research credit for participation. Three participants were excluded because they indicated that they were already native speakers of the chosen language or because they skipped parts of the experiment. A total of 27 participants (15 typing first and 12 speech first) were used in the final analysis. 67% of the participants did the study in Spanish, 26% of participants did the study in French and 7% of participants did the study in Chinese. No significant differences were found in reported measures between languages.

Hypotheses

Given previous research on foreign language learning and voice-driven system design, we set up two hypotheses to explore possible differences between voice-driven and text-driven conditions. First, (H1) learners will try to look up more words in the speaking condition. The cognitive cost of speaking should be less than typing, so we expect learners will be more willing to look up words. Furthermore, previous research on voice-driven learning tools indicates a preference for generating speech

over text [101]. (H2) Learners will find the speaking version more useful. Speaking is a goal of many learners, so we expect practicing speaking will be seen as more practical.

4.2.3 Measures

Usage

Usage was measured as the number of times a learner spoke a new utterance or typed a new phrase. This is a numerical score that counts the number of interactions.

Usefulness measure

After each video section, learners were asked to report how useful they found the system using a continuous slider (0 to 100).

4.3 Findings

Participants' usage frequencies were analyzed using linear mixed regression model. The independent variables were interface type (speech vs. typing) and condition (speech interface first vs. typing interface first). The interface type x condition interaction term was also entered into the model but was non-significant, $F(1, 25) = 0.498$, $p = .487$, $\eta^2 = .015$. The within-subject effect of interface type was significant, $F(1, 25) = 7.23$, $p = .013$, $\eta^2 = .221$, indicating that participants on

average used the speech interface more ($M = 26.52$, $SD = 17.42$) than the typing interface ($M = 18.44$, $SD = 13.95$). The between-subject effect of condition was also significant, $F(1, 25) = 29.45$, $p < .001$, $\eta^2 = .541$, indicating that when participants used the speech interface first, they interacted (speech and typing combined) with the system more ($M = 67.00$, $SD = 18.29$) than those participants that used the typing interface first ($M = 27.33$, $SD = 19.32$). Pairwise comparisons of simple effects also revealed the same finding. For the speech interface, participants in the speech interface first condition used it more often ($M = 38.75$, $SD = 12.08$) than participants in the typing interface first condition ($M = 16.73$, $SD = 14.77$), p < .001. Similarly, for the typing interface, participants in the speech interface first condition used it more often ($M = 28.25$, $SD = 12.75$) than participants in the typing interface first condition ($M = 10.60$, $SD = 9.23$), p = .001.

No significant differences were found in enjoyment, difficulty or reported number of phrases remembered.

4.3.1 Perceived usefulness and usability

A 2x2 mixed ANOVA, with interface type as the within-subject factor and condition as the between-subject factor was performed on participants' perceptions of usefulness. Participants on average found the speech interface more useful ($M = 27.67$, $SD = 22.96$) than the typing interface ($M = 21.41$, $SD = 20.90$). This difference was marginally significant, $F(1, 25) = 3.51$, $p = .073$, $\eta^2 = .120$. The main effect of condition as well as interaction effect were non-significant.

However descriptively, when asked on a preference scale (1-7), with 1 being strongly prefer typing interface, 4 being neutral, and 7 being strongly prefer speech

interface, most participants found the speech method more useful (63% of participants) than the typing method (15% of participants) and some were neutral (22%).

Although some participants had difficulty with the accuracy of the voice recognition engine (e.g. P6: "sometimes it could not understand what I was trying to say"), many participants found speaking was easier than typing (e.g. P18: "...less cognitive overhead than typing", P13: "...you could just say what you heard [instead of typing]"). Furthermore, it eliminated the need to worry about spelling (P19: "The dictionary method was hard because I didn't know how to spell some phrases so it was easier to repeat them.", P12: "I was struggling with how to spell the words so that distracted me."). Other participants indicated that saying words aloud helped with memorization (e.g. P12: "Saying the words out loud makes me remember them more.", P15: "You can pick up on words more quickly by actually saying them out loud.").

However, some participants preferred the text method because it helped to train spelling (e.g. P5: "[the voice method] did not help me with placing accents on letters as the program did that for me", P7: "you may be able to hear the words being spoken in conversation but you may not know how to spell them when writing or reading."). This finding indicates that learner type is important to consider when choosing between text- and voice-driven systems, and perhaps both methods are necessary for comprehensive learning.

While some participants wanted more feedback on pronunciation (e.g. "I won't know if I pronounce or use these words correctly compared to the method of talking this language with the native speaker.", P10: "Saying it out loud doesn't give you a basis for pronunciation so I was saying them incorrectly.") or blamed their pronunciation for trouble with the speech recognition engine (e.g. P21: "[The

voice interface is] harder if you have terrible pronunciation"), many participants found the system to be helpful for improving pronunciation (e.g. P15: "with the voice learning method you can practice speaking the words and sounding them out which is helpful for conversational Spanish.", P23: "Voice helped me understand accents more than typing").

Furthermore, some participants directly discussed learning pragmatic features with the system (e.g. P22: "[M]uch better than doing it from a book. This way I know the right way to pronounce things and the context I might use the phrases in.", P26: "Voice learning is advantageous to other types of learning because you can hear the emotion in a person's voice. I find that Spanish speakers especially give a lot of clues to what they're saying in the way that they're saying it."). Furthermore, the system could help learners overcome lack of confidence in pronunciation (e.g. P11: "[practicing with the voice interface] will not be embarrassing if I pronounce the words badly.").

4.3.2 Learning with native speaker materials

Using materials designed for native speakers is perceived as difficult by many learners. In the comments, many learners indicated that the speech was difficult to follow. When asked about the materials, 60% of learners indicated that the material was very challenging and learners reported an average of 7/10 points when asked how difficult the material was. For example, P3 wrote "I found it very difficult to follow along because they were talking so fast" and P13 wrote "I may have overestimated my French-speaking ability; but I find it difficult to understand films in which the people are speaking fluidly because of how quickly they talk.".

However, just because this activity is challenging does not mean it is not worth doing. Many participants indicated that the use of authentic materials made the learning more valuable (e.g. P10: "I think the materials designed for native speakers is more challenging but it is more original and I believe that I can use it in related situations.", P22: "...made me feel like I was learning phrases I would actually use."). Furthermore, some participants recognized that practice with native speech can can help with communicative goals (e.g. P13: "[T]his is more real-world applicable. When one speaks to native French speakers; they are not going to enunciate every syllable or speak slowly like we learn in French class."). We also speculate that the tool raises awareness in learners about the thoroughness of their understanding. It is easy for a learner to think that they understood the phrase adequately, but repeating phrases out loud can highlight the parts they missed.

Our findings also indicate that learners had diverging perspective on how to learn from native speaker materials. Learners were told that they did not need to understand everything, but some learners felt uncomfortable with this learning method (e.g. P18: "Not as structured. I don't learn the exact grammar. I missed a lot of the dialogue.", P6: "It made it a lot harder because they were talking so fast and it was assumed that I could understand when I really had no idea what was going on"). However, others found this method to be a refreshing change from classroom learning (e.g. P19: "It felt less intimidating because I knew I wasn't supposed to understand everything. In a classroom; a lot of the material is designed for people at your level to there is more pressure to know exactly what everything means.", P21: "...It was also fun to try and repeat the words and trying to see if the translation was correct or made sense."). Given different perspectives on learning through authentic materials, future work should more closely examine these perspectives and could explore designs to promote or support

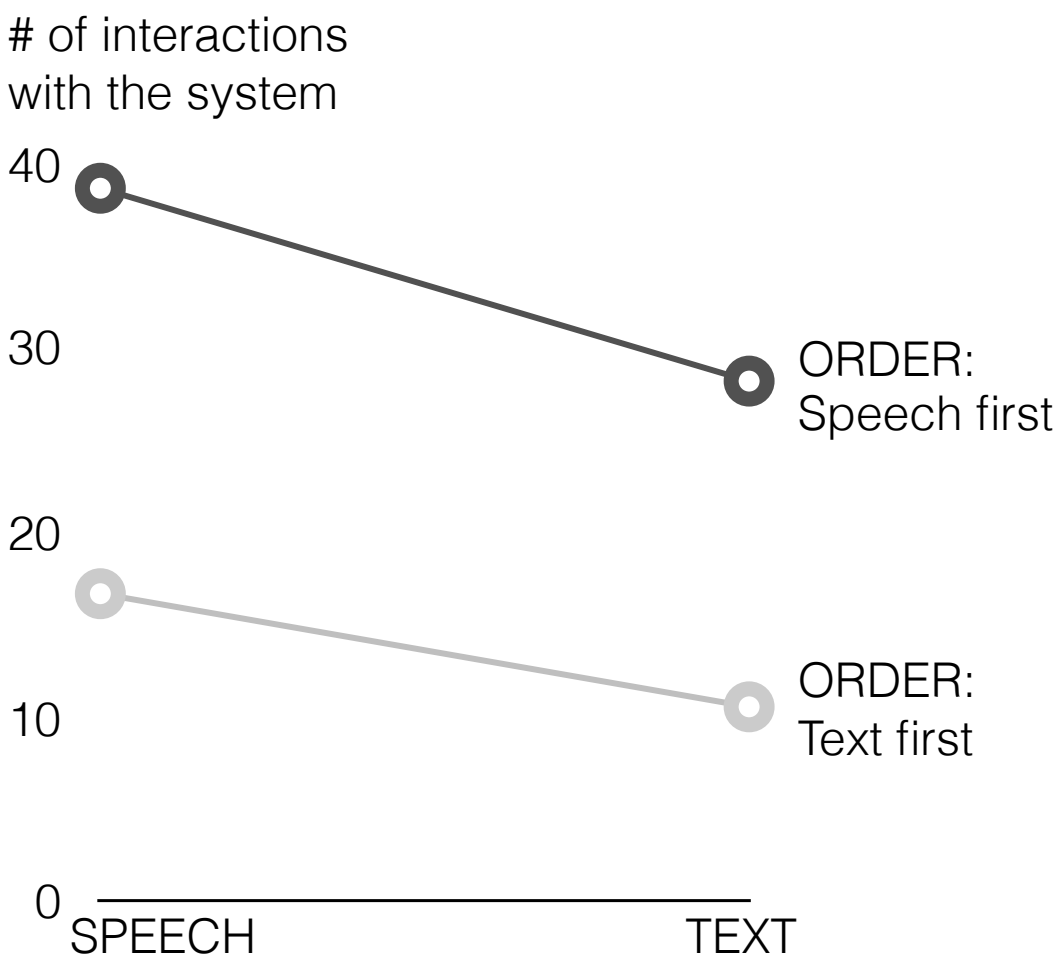

Figure 4.6: Number of interactions with the system for each interface in each ordering. Learners used the speech interface significantly more than the text interface, and using speech first resulted in more overall interactions.

different learning styles.

4.3.3 Order effect

The linear mixed regression analysis reported above revealed that when participants used the speech interface first, they interacted with the system more through

either speaking or typing than those participants that used the typing interface first as shown in Figure 4.6. We found this to be very interesting, but can only speculate on the reasons for this.

We speculate that the speech activity increased learner engagement or sharpened learners' listening ability. In open-ended comments, of participants that started with text, 40% used the word "fast", "rapid" or "quick" to describe the speech, whereas of those with that started with speech only 25% used one of those words to describe the speech. This could indicate that the speech activity prepared learners for better listening. However, more work needs to be done to explore this area. Regardless of the reason, we find the increased engagement with the video after using the voice interface to be an encouraging sign.

4.4 Classroom study

Following the lab study, we set out to expand the design Seiyuu-Seiyuu to better support learning in classrooms. To do this we collaborated with a Japanese instructor at Cornell University (Instructor A) to iteratively improve the design of the website.

4.4.1 Classroom priorities

Because the original system was designed for independent learners, adapting the system for classroom students required some changes. For example, while the original system was entirely open-ended (learners choose videos to learn from and what content to learn in those videos), the teacher we worked with wanted to have more

control over the content. Though we also wished to explore student perspectives on open-ended learning, our initial efforts showed that finding the right balance between open-ended learning activities and teacher-structured activities would be important.

For Instructor A, one of the most enticing aspects of the system is the ability to quickly design assignments with new videos. This helps to keep content up to date and engages students with "young people's language". Using the many new YouTube videos that are uploaded daily (which contain popular Japanese language) teachers can keep assignment content relevant from year to year. We wanted to ensure that any teacher support tools integrated into the system would allow teachers to quickly create assignments around these new videos.

4.4.2 Additional interaction features: targeted repetition and role-playing

Given these priorities, we designed two additional interaction methods for Seiyuu-Seiyuu. The first, targeted repetition, allows instructors to choose language content that they want students to focus on, and the second, roleplaying, allows students to engage in a more open ended way with a video.

Targeted repetition Often instructors have specific content that they want students to learn, either because that material is an important part of the curriculum (e.g. first year students may need to learn many Japanese greetings) or that material may be especially useful for a given student (e.g. young males in Japan may have a specific ways of expressing ideas).

To give teachers control over what content students should focus on within a video, we designed an interaction type we call "targeted repetition". In this assignment type, teachers choose a video and choose lines from the video that students should repeat. As students watch through the video, when one of the target lines is reached, the video will pause after the line and the student will be prompted to repeat that line. Students are allowed to try as many times as they wish. The assignment is completed when the student successfully repeats all of the targeted lines. After the assignment is completed, teachers can listen to each of the student utterances and give feedback offline.

Roleplaying While repetition has been shown to be an effective language learning approach and is helpful when we want learners to focus on specific content, ultimately learners need to comprehend and react to real situations. To better practice this skill, we designed an assignment type where students roleplay as one of the characters in the video. While such a task would be difficult to evaluate automatically, teachers can listen to their students' responses and give feedback afterwards. When creating a roleplaying assignment, teachers choose a video and select lines they want students to roleplay. When a student watches the video, the video will pause just before the target line is reached, and they will be prompted to respond to the situation. Similar to other assignment types, students may record as many times as they wish. When the student is finished recording a line, the video will continue with the audio track replaced with the student's recording. When the assignment is complete, teachers can listen to just the student's utterances or watch the video with student utterances injected. Later, the teacher can share feedback with the student.

4.4.3 Classroom studies

We designed 3 assignments for use in Instructor A's classroom, and ran 2 of these assignments in another Japanese classroom at Cornell University with Instructor B. After completing the assignment, students could optionally complete a survey. We collected surveys during two of the assignments.

4.4.4 Targeted repetition assignment

In the first assignment, student completed a targeted repetition task. This assignment was designed specifically to engage students with "young people's language" or colloquial language used by Japanese youth. Two videos were selected, both from an animated video series called (Paper Rabbit Ropei). Instructor A selected one video that contained speech frequently by young men and another with speech most frequently used by young women. A total of N students completed the assignment and 20 students completed the survey.

4.4.5 Roleplaying and video upload assignment

In a second Japanese classroom (with a different instructor), students did both a roleplaying assignment and uploaded their own video to use with the system. A total of N students completed the assignment and 27 students completed the survey.

4.4.6 Discussion

Student engagement

Students appreciated being able to learn with any video. For example, one student wrote "I like how I can watch any video that I want!" and another student wrote "[I like that] users can import videos they like into Seiyuu-Seiyuu". Another student was excited to learn about the videos their classmates had chosen: "...it offers opportunities for students to share great videos". In total, 9 of 27 students in Instructor B's course mentioned the ability use any video was one of their favorite things about the system. Students also mentioned that these videos had dialogue that is more interesting to them: "The contexts are interesting and are all from animes, TV dramas which I enjoy watching a lot!" In general, interesting content was a major draw for many of the students who used the system.

Similar to the lab study we conducted previously, some students mentioned that they enjoyed speaking aloud and checking if the speech transcription matched their understanding. For example, one student wrote "[I like the] transcript appearing upon voice recognition" and another wrote "The words are going to come up if you say them correctly, which is fun.". However, when the speech recognition did not work, it was frustrating. Several students mentioned that they hoped the recognition quality could be improved.

Helping teachers learning from students

Through the grading interface, teachers can quickly hear students' utterances, and get a sense for issues in pronunciation. They can also view the video with a

student's utterances replacing the original video sound. In open ended tasks, like the roleplay task type, this allows instructors to hear student responses in context and assess whether they make sense or not.

While the bulk of the design effort went into helping teachers assess students, we also found that Seiyuu-Seiyuu helped teachers understand students. While many instructors wish to incorporate native speaker materials such as videos in to their classrooms, it is often difficult to know which materials are relevant and interesting to students. For example, Instructor A mentioned setting up a homework assignment using the animation Gin Tama but found that the assignment quickly became outdated. Each new group of students brings a different set of interests and choosing any individual video will likely alienate some students.

Using Seiyuu-Seiyuu, each learner can choose their own material. This not only ensures that each learner can choose material that is interesting to them, it also allows teachers to get a sense of popular topics and potential new sources of content for a given class. For example, in Instructor A's course, only one student uploaded a live-action drama (rather than an animation) suggesting that animation is more popular for students than live-action. The student upload process also shows new language learning opportunities. For example, some students uploaded videos where manga (Japanese comic books) are read aloud. Instructor A had not encountered these types of videos before and was excited at the possibility of using these types of videos in the future.

Open challenges from the classroom

While many students enjoyed using Seiyuu-Seiyuu and found it useful, there continues to be room for improvement in the system. Similar to the lab study, some

students noted that the system would not always recognize their speech. In order to ensure that student could successfully repeat target phrases, we introduced a system where instructors can train phrases, increasing the likelihood that student utterances will be recognized correctly. However, training phrases is tedious and ideally we would not need teachers to put effort into this. In future iterations we will consider how to reduce the sensitivity of the speech recognizer and while continuing to give feedback on utterance correctness.

Furthermore, learned language was sometimes not relevant. In an assignment for Instructor A's course, a student said phrases like “　” (sea of darkness) and “　” (a machine equal to god”. While interesting to this particular student, these phrases likely would not be used frequently in real conversations. Given freedom, some students might learn particularly eccentric language. While this is not necessarily a drawback, if specific content is important for the student to learn through the exercise, choosing an assignment type that targets phrases (targeted repetition or roleplaying) may be more appropriate.

Expanding audience

We presented the system and results of these studies at the conference for American Council for Teaching of Foreign Languages. Through the conference we were able to get feedback from more teachers and eventually the system was used in Japanese classrooms at other universities.

4.5 Conclusion

The presence of rich context and authentic language makes videos invaluable resources for learning pragmatic competence, but learning with foreign language videos is very difficult. The challenge is to design tools that make the videos more accessible and allow learners to absorb as much as possible from the video materials.

Our results show that using voice is a natural and effective way for learners to engage with videos, and repeating words and phrases from videos can cause learners to engage more with text-based video activities. We found that the tool affords learning through videos that learners enjoy (shown by the variety of cartoons and dramas that learners uploaded during the field study), understanding where phrases might be used (e.g. one participant said: "This way I know the right way to pronounce things and the context I might use the phrases in."), and practicing speech rich in emotion and subtlety (e.g. one participant reflected: "Voice learning is advantageous to other types of learning because you can hear the emotion in a person's voice"). To our knowledge, this is the first study to explore using automatic speech recognition to support video learning, and much work remains to be done in interface design to explore other methods for providing feedback, structuring learning within the system, and boosting learner confidence.

Learning with native speaker materials has the potential to be engaging and effective for learning deep language abilities such as pragmatic competence. This tool is a first step in exploring this space, but much work remains to be done to better understand the what can be learned through native speaker materials, and how best to support learners that wish to use these materials.

CHAPTER 5

A METHOD FOR AUTOMATICALLY ASSESSING LANGUAGE PROFICIENCY USING ELICITED IMITATION WITH TELEVISION PROGRAMS

This section was written in collaboration with Morten Christiansen and Erik Andersen.

Learning from native speaker resources brings a unique set of challenges for language proficiency assessment. In the the grammar-translation approach to language learning, grammar rules are taught one-by-one in a specific order and students learn from lists of words. In this case, assessment is straightforward as tests can simply assess the taught grammar and vocabulary. In the case of the learning with native speaker materials, learners are free to learn what they feel is important from the content, and each learner might use different resources. Therefore, when designing systems for this type of open-ended learning, we should assess general proficiency rather than specific grammar or vocabulary.

However, designers and researchers have few tools available for assessing language proficiency. Self-reported proficiency can be unreliable, and language proficiency tests are difficult to develop and require an extensive time commitment by learners to complete. For example, the national Japanese language proficiency test (the JLPT) requires 3 hours to complete [1] and the national Spanish examination requires around an hour and a half [2]. Furthermore, it is often impossible to use these exams in research or classroom contexts because the exams are only permitted to be administered by specific testing organizations.

[1] jlpt.jp
[2] nationalspanishexam.org

Given these challenges, often researchers and designers use inadequate proficiency measures. For example, in a review of 52 papers on language learning systems from the last 10 years, we find 19 include some measure of language skill, and 15 of those are vocabulary or pronunciation tests. However, language proficiency is not determined by recognizing words or pronunciations in isolation. In real language use scenarios, we hear language in long streams of phrases or sentences, and we do not have time to give attention to each syllable or word separately. Language happens in the here-and-now: we must rapidly recognize and process sounds, words and other units or else heard information is quickly lost [20]. Christiansen and Chater [20] argue that rapid chunking of language input is central to language proficiency. Though there is some work using chunking to measure general skill learning [41, 17, 55], chunking is mostly ignored in learning measures.

Therefore, this chapter seeks to provide a language measure that better assesses learner proficiency and is simple to create and easy to integrate into language studies and systems. We first further explore the concept of language proficiency, and then look at some previous research in methods for evaluating learners. Building on this previous research, we design our language proficiency test. In the test, students listen to utterances in the target language and write down what they can remember of the utterance after they finish listening. This test can easily be constructed from the countless sources of authentic audio on the internet. For example, in our study, we collected audio from Netflix [3] television programs. Where captions are available, the entire test creation process can be automated by selecting target utterances of the desired length in the caption file, and automatically extracting the audio using the times in those captions. Furthermore, the test is very quick. We can learn a lot about a student's proficiency in just a few minutes.

[3] netflix.com

In a study of 97 participants, we show that this measure is well correlated with students' comprehension (measured by having students translate a separate set of heard utterances). Furthermore, we find our measure is better correlated with listening-translation ability than a test derived from a standardized multiple-choice comprehension assessment. Through this work, we hope to offer a better language proficiency measure for designers, researchers and educators.

5.1 Measure design

We set out to construct a measure based on our current understanding of language proficiency and previous research in comprehension assessment.

5.1.1 Defining proficiency

To design our measure, we first carefully consider what we mean by "proficiency". Traditionally language skills have been broken down into components such as grammar, vocabulary, listening, reading, and so on. However, research has shown that our understanding of language is not fully described by grammar or vocabulary in isolation. For example, consider the garden path sentence "The old man the boat". Most listeners will hear two phrases: "the old man" and "the boat". Assuming grammar drives our understanding of language, we would expect that listeners would resolve the sentence to the grammatical interpretation of "the old [N] man [V] the boat [DO]". However, most people do not repair their misunderstandings of garden path sentences [36]. It seems that our propensity to group together frequently co-occurring words better describes how we parse this type of sentence

than grammar.

Work by Christiansen and Chater has shown that when we listen to language, we must rapidly process and chunk incoming language, because our memory for audio is very short and is constantly being overwritten by newly incoming auditory input [20]. This can be easily demonstrated by listening to an audio segment of foreign language and attempting to recall as much of that audio as possible. Most people can only recall a few milliseconds of the audio. Similarly, we can only remember a few units of other types of information. For example, we might have difficulty recalling even the 7 digits of a phone number when spoken aloud. The short duration of short-term memory suggests that a fundamental aspect (perhaps the most important aspect) of language proficiency is our ability to quickly process incoming information and chunk it together into meaningful units. Rapidly chunking language information is central to language proficiency. Although this work originally discussed first language acquisition, the same constraints apply to foreign language learners, so we expect chunking is essential to second language proficiency as well. Furthermore, while at first this constraint may appear to primarily apply to listening and comprehension, further computational modeling work has shown that listening and production are likely part of the same skill [15].

5.1.2 Measuring proficiency

Though language proficiency traditionally has been measured through multiple-choice or true-false questions, these types of measures take a long time to create and are prone to error. Multiple-choice tests specifically have information in the question and answers that can help students strategically to choose the correct answer. For example, in a study comparing students who read a comprehension

passage before answering multiple-choice questions about that passage and students who just read the questions and answers, there was no significant different in the number of correct answers [11]. While it may be possible to construct multiple-choice tests that avoid this issue, these tests require significant effort to build because not only does the test creator need to identify the passage, select a question, and create multiple plausible answers, but also the tests need to be piloted to ensure that the answer cannot be identified using the question and answers alone. Some work has explored automatically generating multiple-choice questions [76], but this approach requires a deep semantic understanding of the text being evaluated. This is especially difficult in non-English languages where data for topic modeling is more scarce, and fewer researchers are exploring these challenges. This combination of issues makes multiple-choice tests less than ideal for assessing learners.

Another commonly used assessment of language ability is cloze [88] test. In all cloze tasks, learners read a passage where some words are deleted, and the learner's task is to fill in the missing words. For example, the learner may see a the sentence "I drove to the _____ to buy some eggs." and need to fill in the word "store". There are many variations in how words should be deleted, for example some tests use randomly selected words, constant-frequency deletions (e.g. every fifth word) or human-selected words based on specific grammar or vocabulary being tested. The difficulty of these tests is sensitive to the deletion method used so it can be difficult to compare between tests [6]. While these tests have been shown to be reliable measures of reading ability, this paradigm would be difficult to adapt to audio contexts. While possible, removing words from audio clips is jarring and would be difficult to accomplish cleanly because of coarticulation between words. Furthermore, it would appear the skills that they test are related to grammar and

vocabulary knowledge [3] which fails to help us get at chunking ability.

While less commonly used, a recall paradigm has more promise as a listening comprehension assessment. Some variations of recall have been studied previously. In one recall test designed to test comprehension, students read a passage and then write down as much information from the passage as they can remember [11]. An instructor identifies "idea units" from the original passage and students' responses are coded for the number of idea units that are included. For example, the "The professor does research on spiders" might have units for "the professor" and "does research". This test gives a comprehensive picture of comprehension, but needs a human to score and has an element of subjectivity. For example, different researchers might choose larger or smaller idea units.

Another variation of recall, elicited imitation, lends itself better to automation and effectively capturing proficiency. Variations of this task have been used to test general statistical learning (e.g. [55]). In elicited imitation tasks, participants listen to utterances and repeat the utterances out loud. Elicited imitation has been shown to be an effective measure of implicit language knowledge in both first and second languages [33, 85, 99]. Erlam [33] suggests that elicited imitation is reconstructive. That is learners must use knowledge about the language in order to complete the task because short-term memory is too short to store all of the information about an utterance. High proficiency learners also correct grammatical mistakes in imitated utterances, suggesting that the meaning of the utterance is remembered rather than the words verbatim [45]. Considering the promise of this method, we used this as a starting point for our measure, viewing it as a natural-language chunking task [15, 20].

5.2 Method

To evaluate whether our measure accurately assessed learner proficiency, we constructed a survey to evaluate learners' comprehension skill (using a translation test), and had learners complete both a multiple-choice comprehension task based on a standardized test and our variation of elicited imitation.

5.2.1 Participants

Data was collected online through Qualtrics survey software and participants were recruited through a university research system. All participants had some experience with Spanish language learning. This could have included high school or college level course or independent learning. Because we wished to understand the effectiveness of our measure across a broad range of skill levels, we allowed participants of any level to participate. Spanish was used as the foreign language because we believed this language would give us the largest and most diverse pool of participants. Data from 97 participants was included in the final analysis (3 participants were excluded from analysis because they had technical issues during the study).

5.2.2 Design

We considered two factors when designing the study. First, we aimed to design the recall measure such that a similar test could easily (either automatically or with minimal researcher input) be created, and we wanted to construct the other measures to create an accurate baseline for learner language proficiency.

To ensure that the test could be easily constructed (and that variants could easily be created), we used online television programs as an audio source. Platforms such as Netflix and YouTube have countless hours of foreign language video, some these videos are already captioned and, if not, captions can easily be added to short segments of the video. Furthermore, the audio is likely more similar to what a learner may hear in real scenarios than a speech synthesizer. When captions are included with the video, it is easy to automatically select utterances for the test (using a combination of utterance length and the vocabulary in the caption) and extract the audio based on the caption times. Audio clips of between 5.5 and 6.5 seconds were selected. We wanted the clips to be long enough to avoid ceiling effects in the data, and, in piloting, we found that this was just beyond the maximum length that most native speakers can remember.

To supplement our Spanish elicited imitation task, we included a debriefing question asking if participants translated during the chunking task or not. This would allow us to assess any differences between participants who completed the task by translating heard utterances into English and then back to Spanish when writing to those who focused on remembering using only Spanish. Furthermore, we included an English elicited imitation task to assess if differences could be explained by general differences in memory or chunking ability rather than language proficiency.

We also included a standard measure of comprehension ability. Multiple-choice questions are frequently used in standardized tests to assess comprehension. However, in measuring foreign language proficiency, multiple-choice questions are often insufficient. A study of multiple-choice questions showed it is difficult to distinguish between participants who read the target passage and those who only read

the questions [11]. Other work has shown that participants generally score better on multiple choice than other types of tests (e.g. cloze or open-ended) [97]. Though we wish to include multiple choice comprehension questions for reference, we wanted to use a more robust and ecologically valid measure as a baseline for comprehension proficiency. Therefore, we designed a translation task to attain the most complete picture of a learner's ability to comprehend language in real-time.

Our translations task is based on the Listening Summary Translation Exam developed by Stansfield et al. [63]. In their task, learners listen to entire conversations in another language and then summarize those conversations into English. This task was developed for FBI workers who would need to summarize conversations as part of their work. We slightly modified the test to better reflect conversational proficiency and better accommodate low-skill learners. In most real language use scenarios, a learner will hear an utterance from another person and need to comprehend and remember the meaning of what was said. Although during communication with a real person a learner would be able to request clarifications, in order to communicate effectively, learners need to comprehend the majority of what is heard the first time. Our translation task was designed to mimic this type of scenario. The learner hears a short audio clip and is asked to translate what was heard. These clips ranged from 12 to 18 words (avg. 15.2 words, 5.81 seconds). By keeping the clips short, we also avoid confounds of memory which have been identified as a potential pitfall of this testing method [4]. Assuming the learner comprehended what was heard, they should be able to summarize this information in a translation even if the translation deviates from the literal meaning of the heard utterance. This type of open-ended translation allows us to estimate how much of a given phrase a student is able to understand.

In sum, our experiment included a recall language test, a set of comprehension multiple choice questions from a standardized test, and an audio translation task. We view the audio translation task as the most ecologically valid and complete assessment of learner's language proficiency. In the analysis, we compare which of the more easily conducted and scored measures (recall and multiple choice) better predict a learner's proficiency.

5.2.3 Procedure

Participants first read a consent form and agreed to take part in the study. Then, participants completed a short exercise to ensure that audio was working on their computer (listen to and audio clip and write down the word "communicate"). Participants then provided some basic demographic information (e.g. "Are you a native speaker of English?", "How would you rate your Spanish proficiency?"). Next participants completed an English recall task (so that we had a baseline for recall ability), and a Spanish recall task.

For both the translation and recall task, participants were instructed to click a button to begin each audio segment. The recall task can also be viewed as a transcription task, as students write down as much as they can exactly as it was said. However, while the audio played, the answer boxes were hidden so that students could not transcribe or translate word-by-word. This was accomplished through Javascript attached to those questions. Once the audio finished playing a single text entry box appeared for the student to supply the translation or transcription along with text instructions (e.g. "Write what you just heard below."). Each audio clip could only be heard once. Participants completed a practice transcription exercise before both the English and Spanish transcription survey sections. In total,

participants completed 5 English transcriptions, 5 Spanish transcriptions and 18 translations from Spanish to English. Following the Spanish transcription task, participants were asked if they were mentally translating during the task.

Next, participants completed 18 questions derived from standardized Spanish tests. This portion of the survey was composed of randomly selected test questions from the The American Association of Teachers of Spanish and Portuguese's National Spanish Exam [75]. Three questions were drawn from each of the 6 levels of the exam (total of 18 questions) in order to build a more general proficiency assessment. Each question had 4 answers and participants were forced to choose an answer. One sentence of context along with the question and answers were displayed to participants. Once participants were ready, they clicked a button to begin the audio clip. These audio clips were 20-40 seconds long. During the audio clip or after it finished playing, participants could choose an answer. Note that unlike the translation and transcription questions, participants did not need to wait until the end to provide an answer. This was in order to better mimic real testing conditions. Participants were could not continue until selecting an answer.

Finally, participants were asked if they had any questions about the study and thanked for their time.

5.3 Analysis

Translations were assessed using both Word Error Rate (WER) and BiLingual Evaluation Understudy (BLEU) [77]. Both metrics have been frequently used to evaluate quality of machine translation and speech recognition systems (e.g. [34]). Both metrics require a set of gold standard translations or transcripts to compare

against. For the elicited imitation task, we already had the gold standard transcriptions from the video captions. These were verified by a native Spanish speaker for accuracy. However, because there are many potential correct translations, and we were interested in meaning rather than precision, we had to take a different approach for creating gold standard translations.

5.3.1 Gold standard translations

Because the goal of the translation task was to assess learners' comprehension of the heard utterances, we did not want the score to reflect naturalness or correctness of the English translation. Therefore we generated gold standard translation using learners' translations as a starting point. Two bilingual speakers of English and Spanish transformed each participant translation into a gold standard translation while minimizing the changes made. For example, for the Spanish phrase "es necesario decirlo de frente" ("I have to be very direct"), a learner might produce "it is necessary to say" which was transformed into "it is necessary to speak directly". This way we ensure that either translation score (WER or BLEU) accurately reflects how much the participant understood rather than other qualities of good translations.

The translators were told to create a correct translation out of each learner translation by adding and removing as few words as possible. Discrepancies between translators were resolved in two stages. First, the guidelines were reviewed, and both of the translators were shown the two translations and asked to produce a new translation based on these that most closely followed the original guidelines. In total, 421 of the 1746 (24%) translations were reevaluated by the translators. In the second round, remaining discrepancies were discussed and a final gold standard

Pearson Correlations ▼

		Translation – BLEU	Translations – WER	Recall – BLEU	Recall – WER	Multiple choice	Self-report
Translation – BLEU	Pearson's r	—	0.981	0.854	0.854	0.539	0.743
	p-value	—	< .001	< .001	< .001	< .001	< .001
Translations – WER	Pearson's r		—	0.811	0.866	0.566	0.753
	p-value		—	< .001	< .001	< .001	< .001
Recall – BLEU	Pearson's r			—	0.749	0.440	0.722
	p-value			—	< .001	< .001	< .001
Recall – WER	Pearson's r				—	0.623	0.727
	p-value				—	< .001	< .001
Multiple choice	Pearson's r					—	0.491
	p-value					—	< .001
Self-report	Pearson's r						—
	p-value						—

Figure 5.1: Correlations between proficiency measures.

translation was agreed upon by the translators. In total, 87 of the 1746 (5%) of the translations were discussed.

5.3.2 WER

Word Error Rate is calculated as: deletions + insertions + substitutions / words in reference A deletion is a missing word when compared to the reference (e.g. "The cat runs", ref: "The black cat runs"), an insertion is an additional word when compared to the reference (e.g "The brown cat runs", ref: "The cat runs"), and a substitution is a changed word (e.g. "The brown cat runs", ref: "The black cat runs"). In order to facilitate comparisons with BLEU scores, statistics and plots use one minus WER rather than the WER score directly. Furthermore, the scores are summed across all trials. Thus, for recall, the possible scores range from 0 to 5 with 5 being a perfect score, and, for translation, the scores range from 0 to 18 with 18 being a perfect score.

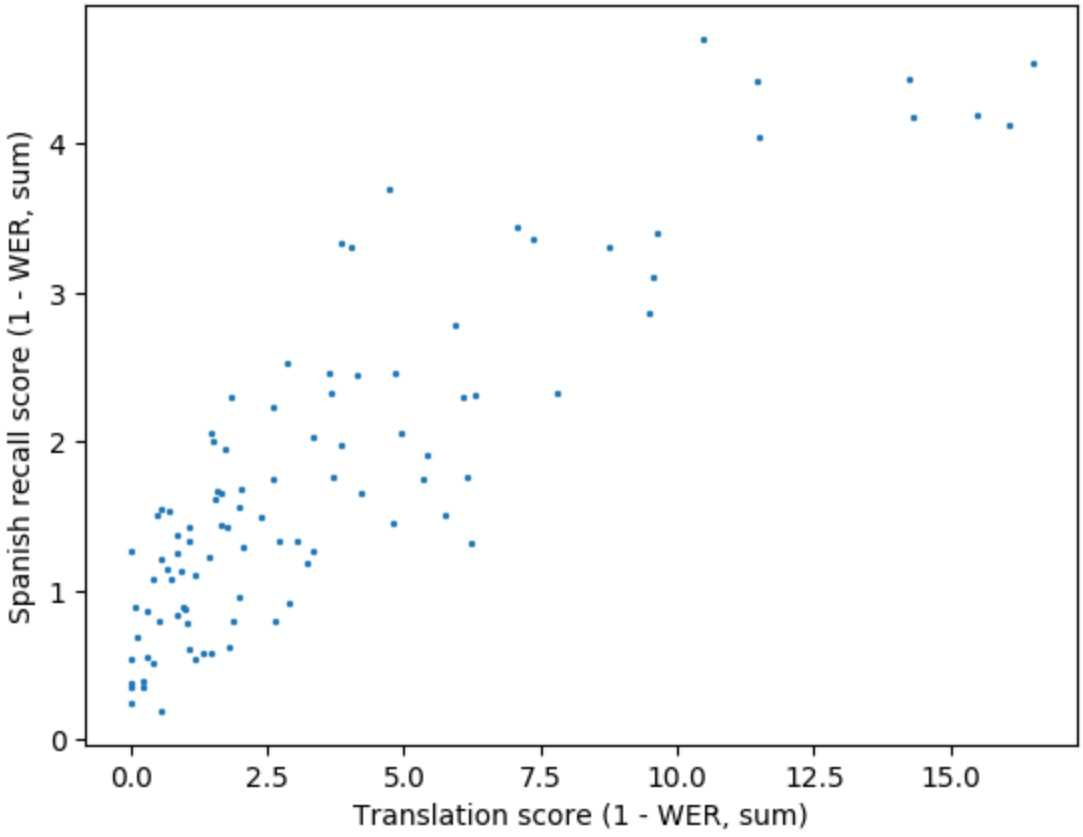

Figure 5.2: This plot shows translation word accuracy using the best reference translation and recall word accuracy (there is only one reference for the recall task.

5.3.3 BLEU

BLEU was designed to take the place of human judgement in evaluation of translation systems [77]. BLEU has been shown to correlate with bilingual judgements of translation quality. The specifics of the method can be found in [77]. Similar to the WER score, the scores are summed across all trials and the score for each trial ranges from 0 to 1. Thus as before, for recall, the possible scores range from 0 to 5 with 5 being a perfect score, and, for translation, the scores range from 0 to 18 with 18 being a perfect score.

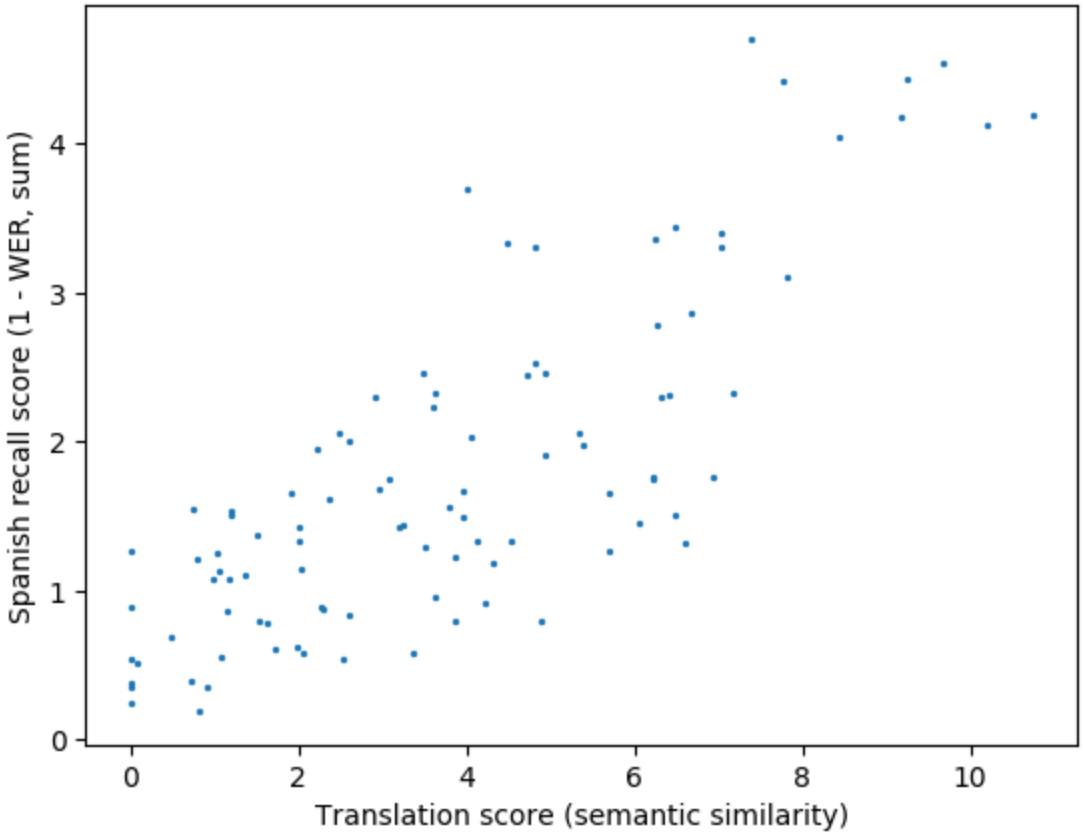

Figure 5.3: This plot shows the semantic similarity score (the semantic similarity between the learners provided translations and the closest reference) and recall word accuracy score.

5.3.4 Semantic similarity

To measure the quality of translations another way, a semantic similarity measure was used to measure the similarity between participant generated translations and the English captions. Because we were most interested in comprehension during the translation task and semantic measures would focus on content rather than English writing skills, this measure is a useful alternative to WER and BLEU. Semantic similarity scores were calculated using the API provided by Han et al. [46]. Their method combines output from a thesaurus (WordNet [71]) and corpus

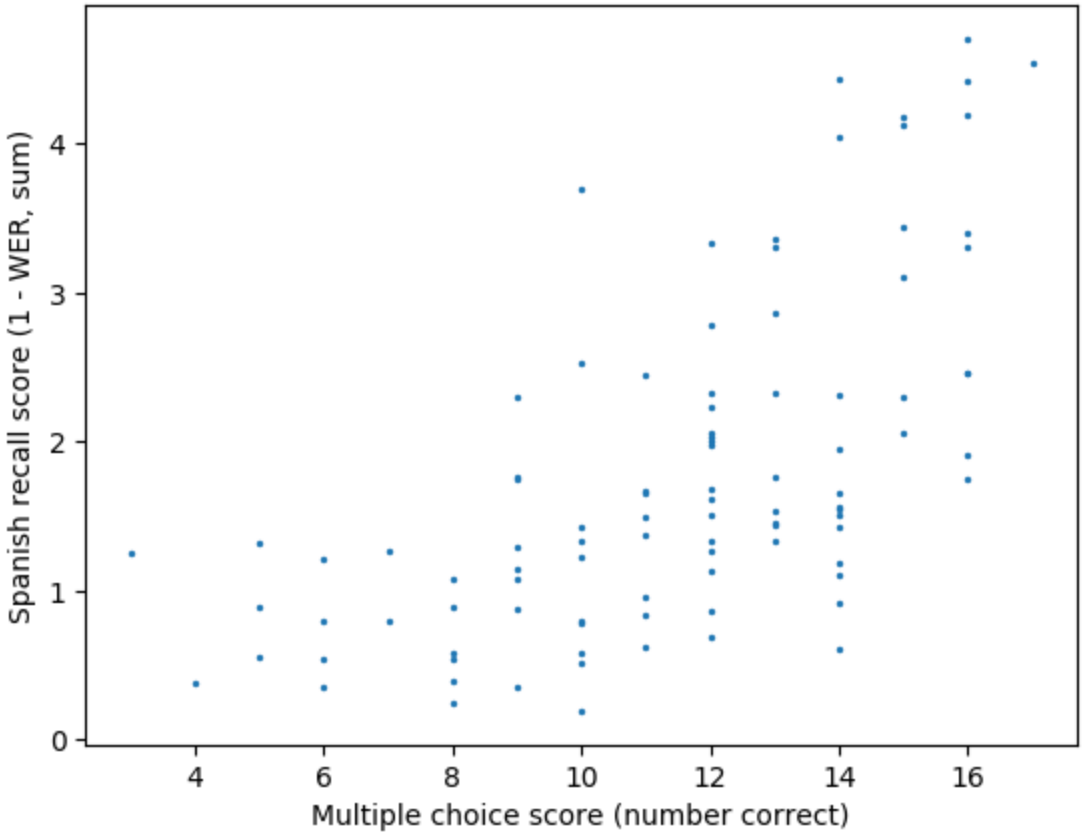

Figure 5.4: This plot shows the multiple choice score (total correct of 18 questions) and recall word accuracy score.

analysis (Web corpus from Stanford WebBase project [87]) in order to calculate semantic similarity.

5.3.5 Correlations

Correlations were calculated between each of the proficiency measures. A table of results are presented in Figure 5.1 and plots are shown in Figures 5.2, 5.3, 5.4, 5.4, and 5.5. Translation assessment using WER and BLEU were found to be highly correlated ($R = 0.981$, $p < .001$), however, using BLEU scoring resulted in many

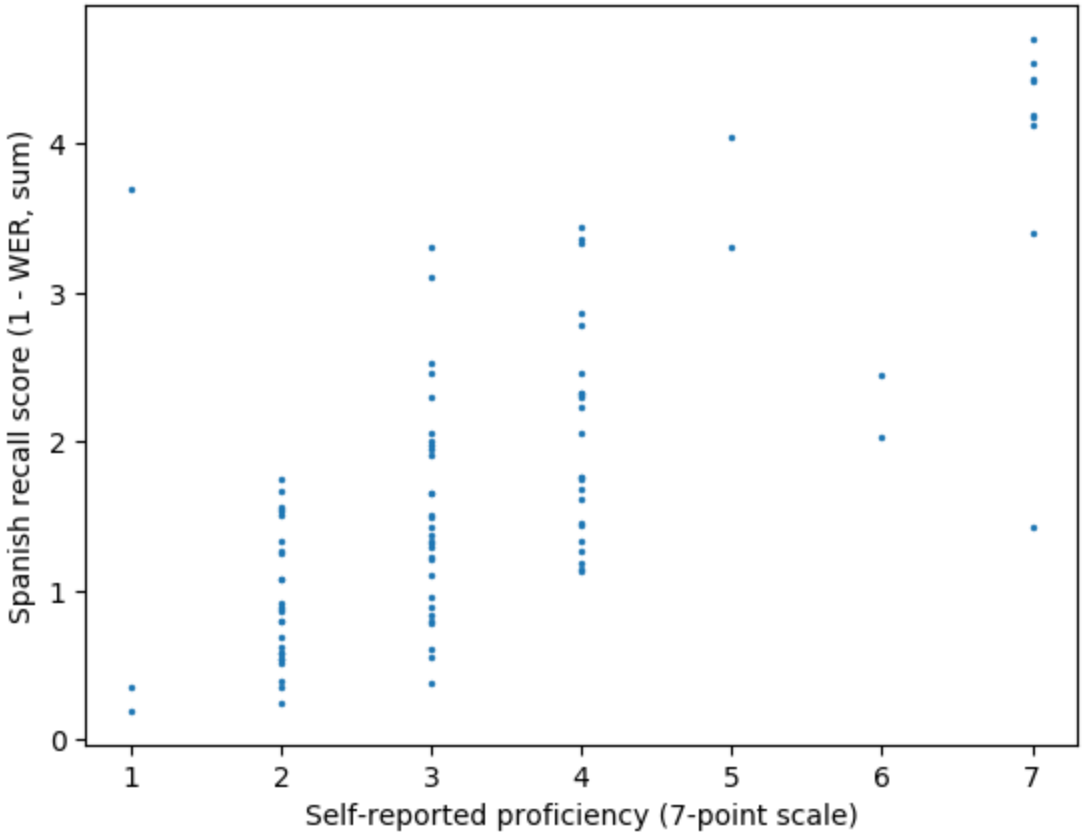

Figure 5.5: This plot shows the learner's self-reported Spanish proficiency (on a scale from 1 to 7 based on the ACTFL proficiency guidelines [2]) and recall word accuracy score.

participants having a score of 0. Therefore, for most the plots, we show only WER scores. Furthermore, although there is a moderate correlation between translation WER scores and multiple-choice scores ($R = 0.539$, $p < .001$), the correlation is significantly higher between translation WER scores and recall WER scores ($R = 0.866$, $p < .001$). This suggests we can better predict a learner's ability to translate using the recall score than the multiple-choice score.

5.3.6 Partial correlations

A partial correlation was calculated to assess whether general memory ability (as measured by English recall) played a significant role in a participant's Spanish recall ability. If the Spanish recall measure is significantly affected by general memory, this would impair the effectiveness of recall as a proficiency measure. However, English recall accounted for only a small part of the overall correlation (R part = 0.114) in a model predicting translation score using Spanish recall and English recall. This gives us confidence that this measure is mostly independent of general memory.

Another partial correlation was calculated to assess whether clip length made a significant difference in the number of words recalled. Although the duration of the audio clips differed by at most 1 second, it is important to understand whether or not the measure is sensitive to variations in audio clip length. The partial correlation was found to be small (R part = 0.155) in a model predicting translation score using recall for an individual audio clip and the duration of that audio clip in seconds. This suggests that small differences in audio clip duration will not affect the test outcome using the recall measure.

5.3.7 High and low proficiency learners

To assess whether the assessment is better suited for low or high proficiency learners, the data was split on the median of the translation score and the lower and upper quartiles were analyzed independently. Correlation was found to be greater in the upper quartiles ($R = 0.814$, $p < .001$) than in the lower quartiles ($R = 0.556$, $p < .001$). This suggests that the test may be less suitable for distinguishing be-

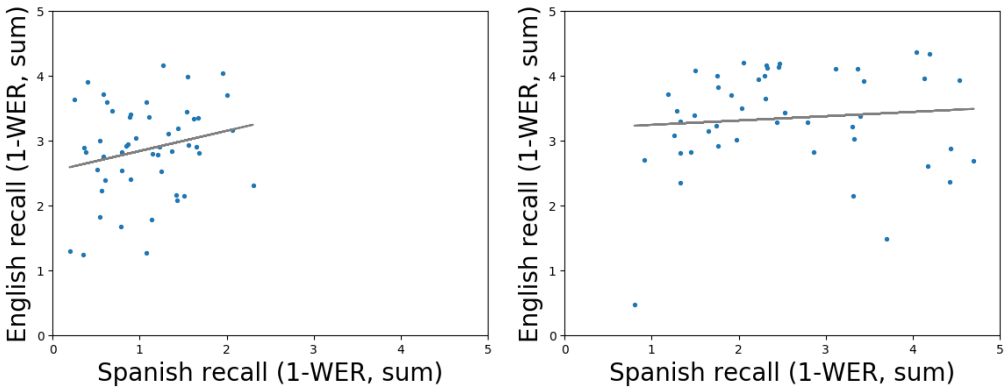

Figure 5.6: Spanish recall against English recall for the lower quartiles (left) and the upper quartiles (right). English recall (a proxy for general memory ability) is more strongly correlated with Spanish recall in the lower quartiles suggesting less skilled learners may be using general memory skills rather than language specific skills at lower proficiencies.

tween skill differences of beginners than distinguishing between high and low proficiency learners or identifying skill differences between high proficiency learners. The weaker correlations at lower learner proficiencies may be due to the measure being more strongly affected by general memory skills for those participants with weaker language skills. Correlation between English recall and WER translation score was found to be stronger in the lower quartiles ($R = 0.224$, $p < .117$) than in the upper quartiles ($R = 0.093$, $p = .533$). However, these correlations were not significant.

5.4 Discussion

The results show that the audio recall task not only accurately reflects learners' ability to comprehend (as assessed through translation), but is actually better at predicting learners' translation ability than the standardized comprehension test.

5.4.1 Improvement over multiple-choice

Our findings support previous work that shows multiple-choice comprehension questions taken from a standardized test may be weak measures of proficiency. In our study, self-reported proficiency level was better correlated with participants' comprehension than the multiple-choice score. While this might be a property of the specific set of multiple-choice questions chosen, it is worth noting that a different set of multiple-choice questions was used during piloting and suggested similar results.

We further note that, as previously discussed, there are a number of drawbacks to using multiple-choice questions, especially in the context of online learning. In our analysis, we noticed that information that can be learned about the student is sparse when using multiple-choice questions. We can only gain a single binary correctness signal for each question answered (the student answered the question correctly or they did not). This means we would likely need many multiple-choice questions to be certain of the student's proficiency, and, combined with the difficulty of creation, compounds the challenge of creating effective multiple-choice tests. This also fails to utilize the student's time effectively. The student must spend significantly more time listening and responding to questions when multiple-choice questions are used because they not only need to listen to the prompt, but also the question and each answer. Although multiple-choice questions continue to be used in the majority of the proficiency tests we looked at. We suggest that we can make better use of students and teachers time, and increase the reliability of our tests by moving away from multiple-choice test structure.

5.4.2 The relationship between translation and recall

This work suggests a strong link between our ability to chunk incoming language and our ability to comprehend. Previous work has shown that we are unable to recall information that is not chunked, and the present study shows that a similar case is true for translation.

Moreover, our findings show that rote repetition cannot describe participants' responses to the elicited imitation task. Low proficiency participants could only recall a few words or no words at all from the target utterance. Though the difference is small, comprehension is more strongly correlated with the absolute number of words recalled than the percent of words recalled. This is likely because limitations on memory are more close related the volume of content being retained that the duration of the content.

5.4.3 Implications for design of language tests

We have shown that a simple recall test based on readily available television programs and subtitles can be used to design reliable foreign language learning proficiency tests. While some care needs to be taken to avoid particularly noisy clips, in general audio tracks from videos are designed to be understood, so most of the audio can be used.

In our elicited imitation test, we used audio clips that we knew most participants would not be able to fully recall. No participant was able to recall every word. While this makes the task more difficult for participants, it ensures that we do have ceiling affects which we saw some evidence of in the multiple choice

test. Previous work has shown that very short utterances can be recalled with rote memorization [33], so we suggest those wishing to use this measure should generally choose longer target phrases.

In future work, it would be useful to evaluate how many recall utterances are needed, the optimal length of clips, and other properties that affect the difficulty of clips. Furthermore, this test will not provide an absolute measure of proficiency, only relative proficiency between learners. Future work could look further at the meaning of the recall score in a more absolute sense to improve comparison between studies. In our study, the length of the clip did not affect the number of words recalled suggesting that scores could be comparable between audio clips provided those audio clips are long enough to avoid ceiling effects.

5.5 Conclusion

In this work we have discussed a method for assessing foreign language proficiency using a quick test that is easily constructed, varied and scored. We hope that by using such a test, we can accelerate the development of effective foreign language learning systems that consider a learner's processing ability instead just grammar and vocabulary tests which fail to provide a complete picture of language proficiency. Furthermore, we have shown that native speaker materials can be used not only as learning resources, but can also as assessment tools. By using these resources in assessments, learners are assessed with tasks that more closely resemble the tasks where learners will eventually exercise their language skills.

CHAPTER 6

CONCLUSION

In the previous 3 chapters, I have illustrated how we can design systems for advanced language proficiency. To summarize, I have shown how we can fulfill key education functions using native speaker materials. Specifically, we can:

- Develop learning materials by using learners to improve the quality of less-than-perfect language learning resources.
- Design learning interactions that support practicing important skills such as speaking in context by leveraging speech recognition and videos.
- Assess learning using a test that is quick for learners, and easy to construct and adaptable for researchers and designers.

Learning content, learning activities and assessment make up the core of any educational curriculum, and we can fulfill all of these functions using native-speaker media, automation, and communicative-learning approaches. Using this framework, we can design systems that scale to even advanced foreign language learners.

This work makes contributions in areas of design, language education, language research methodology, and language learning theory. In design, this work shows how we can build effective language learning experiences around existing materials. In language education, this work generated new learning systems which have been used by independent learners and in classrooms. In language research methodology, this work contributes a new way for researchers to assess learner proficiency using a quick test generated from existing materials. Finally, in language learning theory, this work shows a paradigm shift from the grammar-translation approach to a communicative approach in language system design.

In this chapter, I first discuss a final project which illustrates how these three components can be brought together synergistically in a single system, and then conclude by summarizing the contributions of this work.

6.1 Future work: Combining contextualized learning experiences, learner generated content and proficiency assessment

While taken individually the projects from the previous three chapters show the potential for design around authentic materials, I suggest that approaches from each of these projects can be joined to create scalable and engaging language learning systems. To illustrate how this might be accomplished, I describe final system which brings together elements from each of these areas and applies these ideas to a new type of resource, digital games.

Popular games can have millions of players (e.g. Final Fantasy XIV discussed previously has over 10 million players [1]) and can be tremendously motivating. Furthermore, many games support multiple languages and have many hours of situated, voiced content. For example, in Japanese, a single player roleplaying game Trails in the Sky contains over one million words in the scripts, and contains over 10,000 unique words (about as much as a modern novel). Another game, this time an online roleplaying game, Final Fantasy XIV, contains over 3 millions words in its script and contains over 25,000 unique words. Even supporting learning through a handful of games could help learners move far beyond the content available in most language courses or applications. Given this opportunity, this project explores how we can design for learning with games.

[1] gamespot.com/articles/final-fantasy-14-crosses-10-million-players/1100-6452413/

The system was designed and tested with Japanese, though it could be easily adapted to other languages. Learning Japanese with games brings a number of challenges.

First, Japanese makes use of Chinese characters or Kanji. The characters are logograms, or characters that represent ideas rather than sound. The Japanese kanji do not indicate pronunciation and a single kanji can often have many readings. There are around 2000 standard kanji that students need to learn to pass the highest Japanese proficiency test and there are others beyond that list that may appear in native speaker materials. This means that simply reading text from a game phonetically is challenging. This becomes especially difficult in game contexts where copying and pasting are generally not possible, so looking up readings in a dictionary can be incredibly tedious.

Second, games contain many unique words. Some of these may be specific to a given game's context. For example, words related to chivalry might be absent from most curricula, but common in a game context. While it may not be important that learners internalize this vocabulary, it will hurt their ability to understand the content if they have no way to easily look up this information. Similar to readings, if a learner wishes to look up a word, they would either need to know how to type the word (which is unlikely in the case of rare words) or use a more tedious dictionary lookup method (e.g. hand draw the character or look up the word by radical [2]).

Third, seeing a word, phrase or other language pattern just once may be inadequate to remember that word or phrase over the long term. Repetition over the course of many days or weeks is very important to internalizing language con-

[2]Radicals are parts of characters that can be used to organize them. Paper Japanese dictionaries often include a way to look up characters using one of the 214 radicals.

tent [14]. However, while repetition over a long period of time can help lead to learning, the specific nature of what is being learned should be considered. As previously discussed, our knowledge of language extends beyond the literal meaning or translations to visual contexts (e.g. [92]) and surrounding words (e.g. [20]). Therefore, if learners simply memorize vocabulary lists generated from game texts, they will miss the opportunity to internalize links between the verbal and visual game contexts and the content being learned.

Finally, learning from games is a massive undertaking, and it can be hard to measure progress and know which methods are most effective in the long term. As previously discussed, simply measuring grammar and vocabulary knowledge is insufficient for developing understanding learner proficiency. For this type of learning, there are many learning exercises that could be designed around the game content. For example, learners could complete kanji reading tests, practice speaking the text aloud, practice listening to game audio or practice using the word in new contexts. If we use an ineffective measure of proficiency, we might reach the wrong conclusion about which methods are most effective. For example, it's likely that if we chose a vocabulary test, simply practicing recalling the definitions and readings of words would be the most effective training method. Therefore, our measures need to reflect a deeper understanding of proficiency.

To address these challenges, the system includes four main components: (1) real-time text extraction and kanji reading annotations to help learners read unfamiliar words, (2) a mouseover dictionary for understanding new vocabulary, (3) one-click contextualized learning exercise creation, and (4) automated daily proficiency assessment to help the learner track progress and identify effective methods.

The system functions by scanning system memory for Japanese text and send-

ing any new text that appears in system memory to a web browser window for display and annotation. This scanning approach is generalizable. While different games may use different text encodings, text will always need to be loaded into memory before being displayed on screen. By leveraging a separate rendering system that uses a web browser for rendering, it is easy use the system in many different ways and adapt new to new uses. For example, in the current setup of the system, the annotated text output can be displayed on the same device in a separate window, or the user can play the game on a computer, and view the annotations on a cell phone. A history of output text is maintained, because often context from previous utterances can be helpful in understanding the current utterance.

Similar to previous work in learner edited captions, in this system we can generate a partially correct set of language resources (in this case phonetic annotations for kanji) and learners can improve the annotations as they use the system. The system will automatically append phonetic readings to all of the kanji from the game and display them above the kanji (e.g. Figure 6.1). When the automatic annotation has mistakes, learners can correct them by typing the correct kanji reading. For example, in Figure 6.2, listening to the audio of the game, we know that should be read as “ ” although it has been automatically annotated as “ ”, and the user can enter this correct reading through the system.

A mouseover dictionary is available for helping the learner understand new vocabulary. By hovering over words in the text output window, learners can quickly see multiple possible definitions of a word without needing to switch contexts to a dictionary (Figure 6.3). This system is based on the open source web browser plugin Rikai-kun. [3]. Although this particular tool was selected for the system, many

[3]github.com/melink14/rikaikun

Figure 6.1: Often Japanese Kanji can be difficult to read for learners. To facilitate learning with a game, text from the game is sent to a web interface where the text is annotated with phonetic readings (furigana).

similar tools exist to help learners engage with Japanese web content. Because the interface runs in a web browser, any of these tools could be used with the game text.

While the reading annotation and mouseover dictionary help to understand the game as it is being played, additional work is needed to help the learner retain this information. Spaced repetition has been shown to be one of the most effective ways to retain information over long periods of time [29] and has been used in various apps such as Anki [4] and WanaKana [5]. While often spaced repetition is associated with rote memorization, with games where language content is embedded in rich narrative and visual contexts we have the opportunity to integrate context into spaced repetition. For example, in the case of word learning, we can include the

[4] ankiweb.net
[5] wanakana.com

Figure 6.2: Phonetic annotations are generated automatically and sometimes contain errors. Those errors can be corrected by the learner.

sentence text, audio and a screenshot from the game (Figure 6.4). In the system, learners can hover and click on words in the interface to automatically add this content to the review system. Content in the reviewing system can be reviewed using a simple spaced repetition schedule where a correct answer means the interval of review is doubled and an incorrect answer means means the interval is reset to one day. For example, the first time content is reviewed correctly, it will be shown again after 1 day. If it is reviewed correctly again it will be shown after 2 days, then 4 days, then 8 days and so on. If at any point the learner is unsuccessful in reviewing some content, the interval for that content will be reset to 1 day. This reviewing system ensure that learners can internalize any language content of interest from the game.

Finally, although spaced repetition is effective for scheduling, there are many possibilities for constructing review exercises using the content we have available.

Figure 6.3: Games frequently contain many rare words. To help learners understand the game content, learners can hover the mouse over words in the interface to quickly see definitions.

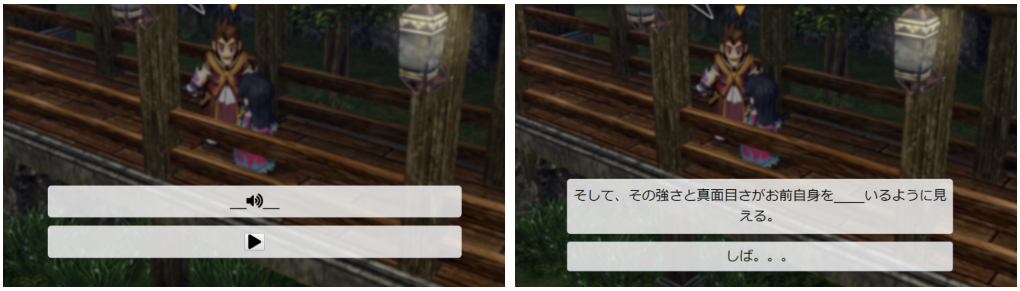

Figure 6.4: Although phonetic readings and definitions can help learners understand content as it is presented, additional effort is needed to retain that information. Using content from the game including text, audio and screenshots, learners can review material later without losing important context.

For example, the system made use of typing reviews (e.g. the learner needs to type a word or phrase), speaking reviews (e.g. the learner needs to speak a word or phrase), and listening reviews (the learner hears the game audio and then needs to type a word or phrase). The system allows for experimentation with different

methods and shows metrics indicating to the learner which methods are most effective for them. The user has the option to enable experiments with different learning exercises, and can track which exercises lead to the best long term retention and transfer. Furthermore, the large repository of audio included with many games (e.g. the game pictured, Trails in the Sky, contains over 30,000 voice clips with approximately 10,000 containing more than 10 words) offers an opportunity to construct proficiency tests as discussed in Chapter 5. This allows for learners to keep track of their overall language proficiency over time and make judgements about the effectiveness of a given study approach.

Such a system can give learners the ability to make sense of game content, improve annotation quality for future learners, retain language knowledge over time and track which approaches are most useful for them. The system shows how the approaches from the previous three chapters can be combined effectively and applied to a new type of resource.

6.2 Closing remarks

This thesis presented a design approach for foreign language education using authentic foreign language resources such as videos and games. Using evidence from psychology, communication and education literature, I showed the importance of contextualized learning experiences for advanced foreign language learners, and showed that authentic resources can be used to provide such experiences. Although authentic resources can be difficult to learn from directly, I show how we can design to support learning with videos and games designed for native speakers.

In Chapter 3, I showed that learners are able to identify errors in automatic

annotations (such as captions or furigana) and correct those errors while learning in the process. We want to let learners use the resources that are engaging and interesting to them, but it's infeasible to expect that every resource will have professionally created annotations for learners. Through the work on imperfect captions as a learning material, we can see that even very early learners can learn with imperfect materials and improve the materials during the learning process.

In Chapter 4, I showed that we can create rich learning experiences around authentic resources such as videos. Through the design of Seiyuu-Seiyuu, I should that we can build learning exercises using speech-shadowing and roleplaying to engage learners deeply with video contexts. Through a studies spanning a year and a half of classroom use, we show support for the ideas that many learners want to use contexts that are relevant and fit their interests.

In Chapter 5, I showed that we can use videos to create effective learning proficiency assessments. Assessing learners can be very challenging and time consuming, but having having a measure of learner proficiency is important for researchers to understand the effectiveness of a given system and important learners in tracking their own progress. The assessment is construct by extracting audio clips from videos and asking learners to listen and recall the audio clips. The learner's response can be automatically scored. This assessment correlates well learner's realtime comprehension. The test is easy to construct for researchers and quick for learners to take. Furthermore, the test is holistic and does not rely on knowledge of specific vocabulary or grammar knowledge. This combination of factors make this test a powerful tool in assessing learners who learn through authentic materials.

This work has taken initial steps towards designing engaging and effective language learning experience using authentic materials. However, more work could

be done to help learners choose content, develop strategies for engaging with materials, and access new domains of native speaker materials.

BIBLIOGRAPHY

[1] New japanese-language proficiency test guidebook, 2009. Available at www.jlpt.jp/e/reference/pdf/guidebook_s_e.pdf.

[2] Actfl proficiency guidelines, 2012. Available at www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012.

[3] J Charles Alderson. The cloze procedure and proficiency in english as a foreign language. Tesol Quarterly, pages 219–227, 1979.

[4] J Charles Alderson and Jayanti Banerjee. Language testing and assessment (part 2). Language teaching, 35(2):79–113, 2002.

[5] Hamid Allami and Amin Naeimi. A cross-linguistic study of refusals: An analysis of pragmatic competence development in iranian efl learners. Journal of Pragmatics, 43(1):385–406, 2011.

[6] Lyle F Bachman. Performance on cloze tests with fixed-ratio and rational deletions. Tesol Quarterly, 19(3):535–556, 1985.

[7] Harry P Bahrick. Two-phase model for prompted recall. Psychological Review, 77(3):215, 1970.

[8] Yutaka Ohno Yoko Sakane Chikako Shinagawa Banno, Eri and Kyoko Takashiki. Genki: an integrated course in elementary Japanese. The Japan Times, 1999.

[9] Kathleen Bardovi-Harlig and Beverly S Hartford. Congruence in native and nonnative conversations: Status balance in the academic advising session. Language learning, 40(4):467–501, 1990.

[10] Julie A Belz. The role of computer mediation in the instruction and development of l2 pragmatic competence. Annual Review of Applied Linguistics, 27:45, 2007.

[11] Elizabeth B Bernhardt. Testing foreign language reading comprehension: The immediate recall protocol. Die Unterrichtspraxis/Teaching German, 16(1):27–33, 1983.

[12] Lawrence F Bouton. A cross-cultural study of ability to interpret implicatures in english. World Englishes, 7(2):183–196, 1988.

[13] David M Cades, Nicole Werner, Deborah A Boehm-Davis, J Gregory Trafton, and Christopher A Monk. Dealing with interruptions can be complex, but does interruption complexity matter: A mental resources approach to quantifying disruptions. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, volume 52,4, pages 398–402. Sage Publications, 2008.

[14] Carlous Caple. The effects of spaced practice and spaced review on recall and retention using computer assisted instruction. 1996.

[15] Nick Chater, Stewart M McCauley, and Morten H Christiansen. Language as skill: Intertwining comprehension and production. Journal of Memory and Language, 89:244–254, 2016.

[16] Matthew Y Chen. Tone sandhi: Patterns across Chinese dialects, volume 92. Cambridge University Press, 2000.

[17] Morten H Christiansen. Implicit statistical learning: a tale of two literatures. Topics in cognitive science, 2018.

[18] Morten H Christiansen, Joseph Allen, and Mark S Seidenberg. Learning to segment speech using multiple cues: A connectionist model. Language and cognitive processes, 13(2-3):221–268, 1998.

[19] Morten H Christiansen and Inbal Arnon. More than words: The role of multiword sequences in language learning and use. Topics in cognitive science, 9(3):542–551, 2017.

[20] Morten H Christiansen and Nick Chater. The now-or-never bottleneck: A fundamental constraint on language. Behavioral and Brain Sciences, 39, 2016.

[21] Kiel Christianson, Andrew Hollingworth, John F Halliwell, and Fernanda Ferreira. Thematic roles assigned along the garden path linger. Cognitive psychology, 42(4):368–407, 2001.

[22] David Crystal. Dictionary of linguistics and phonetics, volume 30. John Wiley & Sons, 2011.

[23] Gabriel Culbertson, Erik Andersen, Walker White, Daniel Zhang, and Malte Jung. Crystallize: An immersive, collaborative game for second language

learning. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, pages 636–647. ACM, 2016.

[24] Gabriel Culbertson, Solace Shen, Erik Andersen, and Malte Jung. Have your cake and eat it too: Foreign language learning with a crowdsourced video captioning system. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. ACM, 2017.

[25] Gabriel Culbertson, Solace Shen, Malte Jung, and Erik Andersen. Facilitating development of pragmatic competence through a voice-driven video learning interface. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pages 1431–1440. ACM, 2017.

[26] Gabriel Culbertson, Shiyu Wang, Malte Jung, and Erik Andersen. Social situational language learning through an online 3d game. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pages 957–968. ACM, 2016.

[27] David Dearman and Khai Truong. Evaluating the implicit acquisition of second language vocabulary using a live wallpaper. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 1391–1400. ACM, 2012.

[28] P Duff. Repetition in foreign language classroom. Second and foreign language learning through classroom interaction, page 109, 2000.

[29] Hermann Ebbinghaus. Memory: A contribution to experimental psychology. Annals of neurosciences, 20(4):155, 2013.

[30] Darren Edge, Kai-Yin Cheng, Michael Whitney, Yao Qian, Zhijie Yan, and Frank Soong. Tip tap tones: mobile microtraining of mandarin sounds. In Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services, pages 427–430. ACM, 2012.

[31] Darren Edge, Stephen Fitchett, Michael Whitney, and James Landay. Memreflex: adaptive flashcards for mobile microlearning. In Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services, pages 431–440. ACM, 2012.

[32] Darren Edge, Elly Searle, Kevin Chiu, Jing Zhao, and James A Landay. Micromandarin: mobile language learning in context. In Proceedings of the

SIGCHI Conference on Human Factors in Computing Systems, pages 3169–3178. ACM, 2011.

[33] Rosemary Erlam. Elicited imitation as a measure of l2 implicit knowledge: An empirical validation study. Applied linguistics, 27(3):464–491, 2006.

[34] Gunnar Evermann. Minimum word error rate decoding. Cambridge University, UK, pages 45–67, 1999.

[35] Thomas A Farmer, Jennifer B Misyak, and Morten H Christiansen. Individual differences in sentence processing. Cambridge handbook of psycholinguistics, pages 353–364, 2012.

[36] Fernanda Ferreira and John M Henderson. Recovery from misanalyses of garden-path sentences. Journal of Memory and Language, 30(6):725–745, 1991.

[37] Thomas J Garza. Evaluating the use of captioned video materials in advanced foreign language learning. Foreign Language Annals, 24(3):239–258, 1991.

[38] Edward Gibson, Caitlin Tan, Richard Futrell, Kyle Mahowald, Lars Konieczny, Barbara Hemforth, and Evelina Fedorenko. Don't underestimate the benefits of being misunderstood. Psychological science, 28(6):703–712, 2017.

[39] Alex Gilmore. Authentic materials and authenticity in foreign language learning. Language teaching, 40(2):97–118, 2007.

[40] Alex Gilmore. "i prefer not text": Developing japanese learners' communicative competence with authentic materials. Language Learning, 61(3):786–819, 2011.

[41] Fernand Gobet, Peter CR Lane, Steve Croker, Peter CH Cheng, Gary Jones, Iain Oliver, and Julian M Pine. Chunking mechanisms in human learning. Trends in cognitive sciences, 5(6):236–243, 2001.

[42] Marta González-Lloret and Katharine B Nielson. Evaluating tblt: The case of a task-based spanish program. Language Teaching Research, 19(5):525–549, 2015.

[43] Yongqi Gu and Robert Keith Johnson. Vocabulary learning strategies and language learning outcomes. Language learning, 46(4):643–679, 1996.

[44] Kenji Hakuta, Yuko Goto Butler, and Daria Witt. How long does it take english learners to attain proficiency?. 2000.

[45] Else Hamayan, Joel Saegert, and Paul Larudee. Elicited imitation in second language learners. Language and Speech, 20(1):86–97, 1977.

[46] Lushan Han, Abhay L Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. Umbc_ebiquity-core: semantic textual similarity systems. In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, volume 1, pages 44–52, 2013.

[47] Tom Hanaway. 'gamified' language app duolingo finally adds japanese.

[48] William J Hardcastle and Nigel Hewlett. Coarticulation: Theory, data and techniques. Cambridge University Press, 2006.

[49] Joshua K Hartshorne, Joshua B Tenenbaum, and Steven Pinker. A critical period for second language acquisition: evidence from 2/3 million english speakers. Cognition, 177:263–277, 2018.

[50] Sayuri Hayakawa and Boaz Keysar. Using a foreign language reduces mental imagery. Cognition, 173:8–15, 2018.

[51] Carol Herron, Bastien Dubreil, Cathleen Corrie, and Steven P Cole. A classroom investigation: Can video improve intermediate-level french language students' ability to learn about a foreign culture? The Modern Language Journal, 86(1):36–53, 2002.

[52] Carol Herron, Matthew Morris, Teresa Secules, and Lisa Curtis. A comparison study of the effects of video-based versus text-based instruction in the foreign language classroom. French Review, pages 775–795, 1995.

[53] Robert Howland, Sachi Urano, and Junichi Hoshino. Sanjigenjiten: computer assisted language learning system within a 3d game environment. In Advances in Computer Entertainment, pages 262–273. Springer, 2012.

[54] IPEK Hulya. Comparing and contrasting first and second language acquisition: implications for language teachers. English Language Teaching, 2(2):155, 2009.

[55] Erin S Isbilen, Stewart M McCauley, Evan Kidd, and Morten H Christiansen.

Testing statistical learning implicitly: a novel chunk-based measure of statistical learning. In the 39th Annual Conference of the Cognitive Science Society (CogSci 2017), pages 564–569. Cognitive Science Society, 2017.

[56] Gabriele Kasper. Can pragmatic competence be taught. NetWork, 6:105–119, 1997.

[57] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J Guo, Robert C Miller, and Krzysztof Z Gajos. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In Proceedings of the 32nd annual ACM conference on Human factors in computing systems, pages 4017–4026. ACM, 2014.

[58] Katherine D Kinzler, Kathleen H Corriveau, and Paul L Harris. Children's selective trust in native-accented speakers. Developmental science, 14(1):106–111, 2011.

[59] Geza Kovacs and Robert C Miller. Foreign manga reader: learn grammar and pronunciation while reading comics. In Proceedings of the adjunct publication of the 26th annual ACM symposium on User interface software and technology, pages 11–12. ACM, 2013.

[60] Geza Kovacs and Robert C Miller. Smart subtitles for vocabulary learning. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 853–862. ACM, 2014.

[61] Stephen D Krashen. Principles and practice in second language acquisition. New York, 1987.

[62] Jean Lave and Etienne Wenger. Situated learning: Legitimate peripheral participation. Cambridge university press, 1991.

[63] Mary Lee Scott, Charles W Stansfield, and Dorry Mann Kenyon. Examining validity in a performance test: The listening summary translation exam (lste)-spanish version. Language Testing, 13(1):83–109, 1996.

[64] Ina Lekkai. Incidental foreign-language acquisition by children watching subtitled television programs. TOJET : The Turkish Online Journal of Educational Technology, 13(4), 2014.

[65] Shiri Lev-Ari and Boaz Keysar. Why don't we believe non-native speak-

ers? the influence of accent on credibility. Journal of Experimental Social Psychology, 46(6):1093–1096, 2010.

[66] Han Z Li. Grounding and information communication in intercultural and intracultural dyadic discourse. Discourse Processes, 28(3):195–215, 1999.

[67] R. McCrum, W. Cran, and R. MacNeil. The story of English. Elisabeth Sifton books. Viking, 1986.

[68] Harry McGurk and John MacDonald. Hearing lips and seeing voices. Nature, 264(5588):746, 1976.

[69] Paul Meara. Connected words: Word associations and second language vocabulary acquisition, volume 24. John Benjamins Publishing, 2009.

[70] Norbert Michel, John James Cater, and Otmar Varela. Active versus passive teaching styles: An empirical study of student learning outcomes. Human Resource Development Quarterly, 20(4):397–418, 2009.

[71] George A Miller. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41, 1995.

[72] Holger Mitterer and James M McQueen. Foreign subtitles help but native-language subtitles harm foreign speech perception. PloS one, 4(11):e7785, 2009.

[73] Duyen T Nguyen and Susan R Fussell. How did you feel during our conversation?: retrospective analysis of intercultural and same-culture instant messaging conversations. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, pages 117–126. ACM, 2012.

[74] David Nunan. Designing tasks for the communicative classroom. Cambridge University Press, 1989.

[75] The American Association of Teachers of Spanish and Portuguese. National spanish examinations. Available at https://www.nationalspanishexam.org/.

[76] Andreas Papasalouros, Konstantinos Kanaris, and Konstantinos Kotis. Automatic generation of multiple choice questions from domain ontologies. In e-Learning, pages 427–434. Citeseer, 2008.

[77] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics, 2002.

[78] Maribel Montero Perez, Elke Peters, and Piet Desmet. Is less more? effectiveness and perceived usefulness of keyword and full captioned video for l2 listening comprehension. ReCALL, 26(01):21–43, 2014.

[79] Jack Richards. Communicative language teaching today. 01 2006.

[80] Victoria Rodrigo, Stephen Krashen, and Barry Gribbons. The effectiveness of two comprehensible-input approaches to foreign language instruction at the intermediate level. System, 32(1):53–60, 2004.

[81] Teresa Secules, Carol Herron, and Michael Tomasello. The effect of video context on foreign language learning. The Modern Language Journal, 76(4):480–490, 1992.

[82] Kaoru Sekiyama and Yoh'ichi Tohkura. Mcgurk effect in non-english listeners: few visual effects for japanese subjects hearing japanese syllables of high auditory intelligibility. The Journal of the Acoustical Society of America, 90(4):1797–1805, 1991.

[83] Harriet D Semke. Effects of the red pen. Foreign language annals, 17(3):195–202, 1984.

[84] Pamela Sherer and Timothy Shea. Using online video to support student learning and engagement. College Teaching, 59(2):56–59, 2011.

[85] Dan I Slobin and Charles A Welsh. Elicited imitation as a research tool in developmental psycholinguistics. 1967.

[86] Steven M Smith, Arthur Glenberg, and Robert A Bjork. Environmental context and human memory. Memory & Cognition, 6(4):342–353, 1978.

[87] Stanford. Stanford webbase project, 2001.

[88] Wilson L Taylor. "cloze procedure": A new tool for measuring readability. Journalism Bulletin, 30(4):415–433, 1953.

[89] Japanese Language Proficiency Test. N1-n5: Summary of linguistic competence required for each level. Available at http://www.jlpt.jp/e/about/levelsummary.html.

[90] Pavel Trofimovich and Elizabeth Gatbonton. Repetition and focus on form in processing l2 spanish words: Implications for pronunciation instruction. The Modern Language Journal, 90(4):519–535, 2006.

[91] Michio Tsutsui and Masashi Kato. Designing a multimedia feedback tool for the development of oral skills. 2001). CALL-The challenge of Change: Research & Practice, pages 81–88, 2001.

[92] Endel Tulving and Donald M Thomson. Encoding specificity and retrieval processes in episodic memory. Psychological review, 80(5):352, 1973.

[93] Robert Vanderplank. Déjà vu? a decade of research on language laboratories, television and video in language learning. Language teaching, 43(01):1–37, 2010.

[94] Luis von Ahn. Duolingo: learn a language for free while helping to translate the web. In Proceedings of the 2013 international conference on Intelligent user interfaces, pages 1–2. ACM, 2013.

[95] Helen Williams and David Thorne. The value of teletext subtitling as a medium for language learning. System, 28(2):217–228, 2000.

[96] Werner Winiwarter. Mastering japanese through augmented browsing. In Proceedings of International Conference on Information Integration and Web-based Applications & Services, page 179. ACM, 2013.

[97] Darlene F Wolf. A comparison of assessment tasks used to measure fl reading comprehension. The Modern Language Journal, 77(4):473–489, 1993.

[98] Jean Wong. Delayed next turn repair initiation in native/non-native speaker english conversation. Applied Linguistics, 21(2):244–267, 2000.

[99] Shu-Ling Wu and Lourdes Ortega. Measuring global oral proficiency in sla research: A new elicited imitation test of l2 chinese. Foreign Language Annals, 46(4):680–704, 2013.

[100] Liu Xun. New practical Chinese reader, volume 4. Beijing Language and Culture University Publishing House, 2004.

[101] Dongwook Yoon, Nicholas Chen, François Guimbretière, and Abigail Sellen. Richreview: blending ink, speech, and gesture to support collaborative document review. In Proceedings of the 27th annual ACM symposium on User interface software and technology, pages 481–490. ACM, 2014.

[102] Chien Wen Yuan, Leslie D Setlock, Dan Cosley, and Susan R Fussell. Understanding informal communication in multilingual contexts. In Proceedings of the 2013 conference on Computer supported cooperative work, pages 909–922. ACM, 2013.